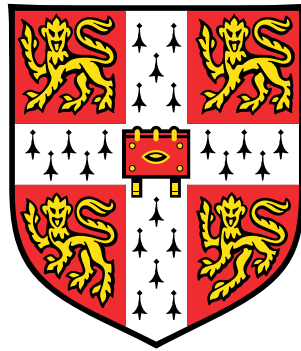


# Exploring Multivariate Gene-Environment Interactions: Models And Applications

**Rachel Moore**

Wellcome Sanger Institute  
Darwin College  
University of Cambridge



This dissertation is submitted for the degree of  
Doctor of Philosophy

November 2018





# Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the Faculty of Biology.



# Abstract

Complex diseases are driven by multiple risk factors, including genetic variants, environmental exposures and interactions between the two. The advent of GWAS in 2005 and subsequent methodological advances have increased our knowledge of the genetic risk factors underpinning complex diseases. In addition, some research exploring genotype-environment interaction ( $G \times E$ ) effects has been conducted, revealing that multiple environments are linked to interaction effects at a single locus for a given trait. However, correlation between these identified environments renders interpretation of the results difficult. This, together with the collation of large-scale biobanks that contain a multitude of phenotypic and environmental data (facilitating an increase in the number of  $G \times E$  effects detected) has generated the need for methods that jointly account for  $G \times E$  at multiple environments. Such methods may also increase the power to detect interaction effects by aggregating modest or weak  $G \times E$  effects across environments and in addition enable additional phenotypic variance to be explained. Thus, the aim of this thesis is to provide suitable methods to identify variants subject to  $G \times E$  effects, jointly accounting for multiple environmental exposures and explore these effects across a range of phenotypes using the UK Biobank data.

In Chapter 2, I describe the structured linear mixed model (StructLMM), a novel computationally efficient multivariate  $G \times E$  framework. This model can be used to test for interaction or association effects. The latter accounts for possible heterogeneity in variant effects across individuals due to differences in environmental exposures, thus enabling the detection of variant effects that might otherwise be masked due to the presence of interaction effects. I show through the use of simulation experiments that StructLMM is robustly calibrated and in general, better powered than existing interaction and association tests.

In Chapter 3, I present an application of StructLMM, where I identify significant interaction effects with 64 lifestyle-based factors for BMI using the UK Biobank data. In addition, I show that the StructLMM association test can be used to identify loci

with genotype-environment contributions. Subsequently, I explore characteristics of loci with significant interaction effects, including the fraction of the genetic variance that is explained by  $G \times E$  and the environmental profiles that increase or decrease phenotypic risk, using methods that are implemented as part of StructLMM.

In Chapter 4, I apply the StructLMM interaction test to multiple cardiometabolic traits using the UK Biobank data, facilitating exploration of shared  $G \times E$  architecture. Additionally, I provide preliminary estimates of the amount of phenotypic variation that can be explained by  $G \times E$  effects, relative to marginal association effects.

Taken together, the work in this thesis demonstrates the need and advantages of jointly modelling interaction effects at multiple environments, providing a new computationally efficient method to achieve this. Combined with the recent and ongoing generation of large biobanks, further research in this field has the potential to advance our understanding of complex traits and diseases.

# Acknowledgements

I would like to thank my supervisors Dr. Inês Barroso and Dr. Oliver Stegle, both of whom have been outstanding mentors. They have provided me the freedom to explore my own ideas, whilst at the same time giving fantastic advice, guidance and encouragement.

I would also like to thank Dr. Francesco Paolo Casale who spent many hours, in particular during the first year of my PhD, introducing and explaining to me concepts within the field of Statistical Genetics. This help and guidance was instrumental in making my PhD productive and exciting. I would also like to thank Dr. Danilo Horta, not only for his recent involvement in finalising the software to accompany the StructLMM method, but in addition for his continued patience in helping me with all things computer-related!

Others I would like to thank are Dr. Na Cai for interesting and useful discussions, Fernando Riveros Mckay Aguilera for helping me access and find my way around the UK Biobank data and Felicity Payne for introducing me to the many useful tools and resources during my rotation project when I first joined the Barroso team. I would also like to thank the other past and present members of the Stegle and Barroso groups for their thoughts and advice, as well as providing a fun environment to work in. The extremely varied interests of different group members and great discussions during lab meetings and group retreats have meant that I have learnt much over the past three years. Finally, I would like to acknowledge the financial support that I have received to carry out this work made available through the Mathematical Genomics and Medicine PhD programme funded by Wellcome.

On a more personal note, I would like to thank my family. In particular, my parents and husband for their continued and unwavering support, encouragement and for being interested (or at least pretending to be interested!) in my research.



# Contents

<b>Declaration</b>	<b>I</b>
<b>Abstract</b>	<b>II</b>
<b>Acknowledgements</b>	<b>IV</b>
<b>List of Tables</b>	<b>VIII</b>
<b>List of Figures</b>	<b>IX</b>
<b>List of abbreviations</b>	<b>XI</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Identifying the genetic basis of traits and diseases . . . . .	1
1.1.1 Mendelian and complex traits and diseases . . . . .	1
1.1.2 Genetic linkage studies . . . . .	2
1.1.3 Common disease/common variant hypothesis . . . . .	3
1.1.4 Characterising the LD structure of the human genome . . . . .	5
1.1.5 Genome-wide association analyses . . . . .	7
1.1.6 Caveats of association analyses . . . . .	10
1.2 Advances in association analyses . . . . .	15
1.2.1 Rare variants . . . . .	15
1.2.2 Increased sample sizes . . . . .	17
1.2.3 Multi-trait methods . . . . .	18
1.2.4 Pleiotropy and PHEWAS . . . . .	19
1.3 What about the other components of complex traits and diseases? . .	21
1.3.1 Other contributing factors . . . . .	21
1.3.2 Epistasis and $G \times E$ . . . . .	22
1.4 Thesis overview . . . . .	26
<b>2 StructLMM: a linear mixed model approach to study multivariate</b>	

<b>genotype-environment interactions</b>	<b>28</b>
2.1 Introduction . . . . .	28
2.2 StructLMM . . . . .	32
2.2.1 The model . . . . .	33
2.2.2 StructLMM interaction and association tests . . . . .	35
2.2.3 Simplifying assumptions . . . . .	38
2.2.4 Environment covariance, $\Sigma$ . . . . .	39
2.2.5 Statistical testing . . . . .	41
2.2.6 Computational complexities . . . . .	53
2.2.7 Relationship to existing methods . . . . .	60
2.3 Comparison partners . . . . .	65
2.3.1 Single environment models . . . . .	66
2.3.2 Multi-environment models . . . . .	68
2.3.3 Linear association models . . . . .	69
2.3.4 Summary of comparison partners . . . . .	71
2.4 Simulation experiments . . . . .	72
2.4.1 Simulation data . . . . .	72
2.4.2 Phenotype simulation procedure . . . . .	74
2.4.3 Simulation approach . . . . .	76
2.4.4 Simulation results . . . . .	78
2.4.5 Simulation results examining the effect of specific environmental properties . . . . .	84
2.5 Summary and discussion . . . . .	91
<b>3 Application of StructLMM to identify genotype-environment interaction effects that influence body mass index in UK Biobank</b>	<b>93</b>
3.1 Introduction . . . . .	93
3.2 Methods to explore identified loci . . . . .	96
3.2.1 Estimating the fraction of the genetic variance driven by $G \times E$	97
3.2.2 Exploration of the environments that drive the $G \times E$ effects .	98
3.2.3 Estimation of per-individual allelic effect sizes due to $G \times E$ . .	100
3.2.4 Estimation of the aggregate environment driving the $G \times E$ effect at a variant . . . . .	101
3.2.5 Computational complexities of methods used to explore loci .	102
3.3 Methods . . . . .	103
3.3.1 UK Biobank data processing . . . . .	103
3.3.2 Calibration . . . . .	105
3.3.3 Defining independent and exclusive loci . . . . .	108



3.4	Results . . . . .	108
3.4.1	Calibration assessment . . . . .	108
3.4.2	Interaction test results . . . . .	110
3.4.3	Genome-wide association results . . . . .	114
3.4.4	Exploration of significant loci . . . . .	119
3.5	Summary and discussion . . . . .	128
<b>4</b>	<b>Exploring genotype-environment interaction effects across different phenotypes</b>	<b>131</b>
4.1	Introduction . . . . .	131
4.2	Methods . . . . .	136
4.2.1	UK Biobank data preprocessing . . . . .	137
4.2.2	Calibration . . . . .	139
4.2.3	Defining loci . . . . .	139
4.2.4	cFDR . . . . .	140
4.2.5	Estimation of phenotypic variance explained . . . . .	141
4.3	Results . . . . .	141
4.3.1	Basal metabolic rate significantly associated variants . . . . .	142
4.3.2	Calibration assessment . . . . .	143
4.3.3	Association of variants with the considered phenotypes . . . . .	143
4.3.4	Interaction of variants for different considered phenotypes . . . . .	145
4.3.5	Distribution of $\rho$ across different traits . . . . .	149
4.3.6	Relative importance of $G \times E$ effects compared to persistent genetic effects . . . . .	151
4.4	Summary and discussion . . . . .	153
<b>5</b>	<b>Concluding remarks</b>	<b>158</b>
	<b>References</b>	<b>163</b>
	<b>Appendix A: Copy of manuscript entitled, ‘A linear mixed-model approach to study multivariate gene-environment interactions’, published by Nature Genetics</b>	<b>199</b>
	<b>Appendix B: Summary of results from StructLMM interaction test described in Chapter 4 for the different considered phenotypes</b>	<b>266</b>



# List of Tables

2.1	Computational complexity of different operations required for the StructLMM interaction and association tests . . . . .	60
2.2	Comparison methods considered . . . . .	71
2.3	Parameters used for simulation experiments . . . . .	76
B1	Loci containing significant interaction effects when examining ten phenotypes . . . . .	266
B2	Overlap of loci containing significant interaction effects for the ten considered phenotypes . . . . .	279



# List of Figures

1.1	Genetic linkage analysis to identify the genetic region linked to EDS-VIII in a Swedish pedigree . . . . .	4
1.2	Identification of tag SNPs . . . . .	6
1.3	Imputation of genotype data . . . . .	8
1.4	Example Manhattan plot . . . . .	10
1.5	Example quantile-quantile (QQ) plots . . . . .	11
1.6	Different relationships between allele dosages, environmental exposures and phenotypes . . . . .	23
2.1	Different $G \times E$ scenarios that can be tested using StructLMM and corresponding examples of environment covariance structures, $\Sigma$ . . .	36
2.2	Visualisation of equivalence of Eq. 2.47 and Eq. 2.48 using the probability density function of $q_{\rho_r}$ . . . . .	47
2.3	Calibration of StructLMM interaction and association test . . . . .	79
2.4	Power comparison of different methods . . . . .	81
2.5	Calibration comparison of different implementations of multi-environment tests . . . . .	82
2.6	Power comparison of different implementations of multi-environment tests . . . . .	83
2.7	Calibration of StructLMM interaction and association tests when environments are simulated to be skewed or binary with rare event frequency . . . . .	85
2.8	Power comparison of different methods examining the effect of using skewed and binary environments . . . . .	87
2.9	Calibration of StructLMM in the presence of heritable environments .	89
2.10	Power comparison of different methods in the presence of heritable environments . . . . .	90
3.1	Covariance structure, $\Sigma$ , of UK Biobank individuals based on 64 environmental variables . . . . .	106

3.2	Correlation of the 64 environments based on 252,188 UK Biobank individuals . . . . .	107
3.3	Calibration of interaction and association tests for BMI on UK Biobank data . . . . .	109
3.4	Local Manhattan plots for the four interaction loci identified by StructLMM-int . . . . .	111
3.5	Comparison of interaction results at the 97 GIANT loci . . . . .	111
3.6	Comparison of interaction results at genome-wide significant variants in UK Biobank . . . . .	113
3.7	Comparison of LMM-Renv and LM results on UK Biobank data . . .	115
3.8	Local Manhattan plots for three association loci exclusively identified by StructLMM in UK Biobank . . . . .	116
3.9	Comparison of StructLMM and LMM-Renv association tests on UK Biobank data at all 7,515,856 tested variants . . . . .	117
3.10	Distribution of the estimated extent of $G \times E$ for significant loci identified by the StructLMM association test and LMM-Renv on UK Biobank data . . . . .	118
3.11	Comparison of multi-environment association tests on UK Biobank data . . . . .	119
3.12	Evidence for each environment contributing to the identified $G \times E$ interaction effects using Bayes factors . . . . .	120
3.13	Exploration of the environments contributing to the identified $G \times E$ interaction effects using Bayes factors . . . . .	122
3.14	Exploration of the aggregate environment effect at identified $G \times E$ loci . . . . .	124
3.15	Out of sample prediction of per-individual allelic effect sizes . . . . .	126
3.16	Rank correlation of per-individual genetic effect sizes across loci for UK Biobank data . . . . .	127
4.1	Manhattan plot of genome-wide association results for BMR . . . . .	142
4.2	Calibration of interaction and association tests on UK Biobank data .	143
4.3	Association results with the considered phenotypes . . . . .	144
4.4	Application of the cFDR to genome-wide interaction results . . . . .	146
4.5	Interaction results with the considered phenotypes . . . . .	148
4.6	Distribution of $\rho$ for the considered phenotypes . . . . .	150
4.7	Estimated fraction of phenotypic variation explained by persistent genetic and $G \times E$ effects . . . . .	152

# List of abbreviations

<b>AUC</b>	Area under the curve
<b>BF</b>	Bayes factors
<b>BMI</b>	Body mass index
<b>BMR</b>	Basal metabolic rate
<b>Body fat %</b>	Body fat percentage
<b>Bp</b>	Base pairs
<b>CM</b>	Centimorgan
<b>CD/CV</b>	Common disease/common variant
<b>CEU</b>	Utah Residents (CEPH) with Northern and Western European Ancestry
<b>CFDR</b>	Conditional false discovery rate
<b>DEXA</b>	Dual energy X-ray absorptionmetry
<b>Df</b>	Degree(s) of freedom
<b>DBP</b>	Diastolic blood pressure
<b>EDS</b>	Ehlers-Danlos syndrome
<b>EMERGE consortium</b>	Electronic medical records and genomics consortium
<b>EMR</b>	Electronic medical records
<b>FDR</b>	False discovery rate
<b>FIN</b>	Finnish in Finland

<b>FWER</b>	Family wise error rate
<b>GBR</b>	British in England and Scotland
<b>G×E</b>	Genotype-environment interaction
<b>GIANT</b>	Genetic investigation of anthropometric traits
<b>GRS</b>	Genetic risk score
<b>GWAS</b>	Genome-wide association study
<b>HDL</b>	High density lipoprotein
<b>HC</b>	Hip circumference
<b>HRC</b>	Haplotype Reference Consortium
<b>HWE</b>	Hardy-Weinberg equilibrium
<b>IBS</b>	Iberian Population in Spain
<b>I.i.d.</b>	Independent and identically distributed
<b>Kb</b>	Kilobase
<b>LD</b>	Linkage disequilibrium
<b>LDSC</b>	LD score regression
<b>LM</b>	Linear model
<b>LML</b>	Log marginal likelihood
<b>LMM</b>	Linear mixed model
<b>LRT</b>	Likelihood ratio test
<b>MAF</b>	Minor allele frequency
<b>Mb</b>	Megabase
<b>MLE</b>	Maximum likelihood estimate
<b>P-adj</b>	Adjusted P value
<b>PAGE</b>	Prenatal assessment of genomes and exomes
<b>PC</b>	Principal component



<b>PCA</b>	Principal component analysis
<b>PEF</b>	Peak expiratory flow
<b>PHEWAS</b>	Phenome-wide association study
<b>PHEWIS</b>	Phenome-wide interaction study
<b>QC</b>	Quality control
<b>QQ</b>	Quantile-quantile
<b>REML</b>	Restricted maximum likelihood
<b>SBP</b>	Systolic blood pressure
<b>SKAT</b>	Sequence kernel association test
<b>SKAT-O</b>	Optimal sequence kernel association test
<b>SNP</b>	Single nucleotide polymorphism
<b>TDI</b>	Townsend deprivation index
<b>TPR</b>	True positive rate
<b>TSI</b>	Toscani in Italia
<b>WC</b>	Waist circumference
<b>WGS</b>	Whole genome sequencing
<b>WES</b>	Whole exome sequencing
<b>WHO</b>	World Health Organisation
<b>WHR</b>	Waist-to-hip ratio
<b>W. r. t.</b>	With respect to
<b>WTCCC</b>	Wellcome Trust Case Control Consortium



# Chapter 1

## Introduction

### 1.1 Identifying the genetic basis of traits and diseases

#### 1.1.1 Mendelian and complex traits and diseases

The genetic basis underpinning traits and diseases was a hotly debated topic in the early 20<sup>th</sup> century<sup>1-8</sup>, commonly referred to as the ‘Biometric-Mendelian debate’. Opinions were divided between Gregor Mendel’s laws of inheritance<sup>9</sup> and those originally put forward by Francis Galton<sup>10</sup>. It is now widely accepted that both modes of inheritance exist, such that some traits are classified as Mendelian and others as complex.

Mendelian traits are also referred to as monogenic disorders, which as the name suggests, are driven by mutations within a single gene. Examples of such traits include X-linked muscular dystrophies, cystic fibrosis, Fanconi anaemia, the classic form of Ehlers-Danlos syndrome and phenylketonuria<sup>11-18</sup>. Phenylketonuria is an example of a monogenic disorder that only manifests under specific environmental conditions; specifically the phenotype only occurs when there are both mutations within the *PAH* gene and when an individual is exposed to phenylalanine (naturally occurring in dietary proteins)<sup>19,20</sup>. Mendelian traits are typically rare in the general population but often cluster in families<sup>21</sup>.

In comparison, complex traits are typically common in the general population<sup>21</sup>. They are driven by a combination of multiple genetic risk factors across the genome (thus they are often referred to as polygenic traits), environmental risk factors, as

well as interaction effects between genetic variants and environmental exposures<sup>21,22</sup>. The combined contribution of the genetic and environmental factors, results in a continuous range of phenotypic values; hence complex traits are sometimes referred to as quantitative traits<sup>23</sup>. Such complex traits include height, weight and blood pressure<sup>23</sup>. However, for some complex traits, the combination of genetic and environmental factors can be viewed as a predisposition measure that when combined with a suitable penetrance (defined here as the probability of having a disease given the predisposition score) mapping function results in an observed binary outcome (i.e. either diseased or not)<sup>24</sup>. Examples of complex traits with binary outcomes include heart disease, type 2 diabetes, schizophrenia, asthma and cancer<sup>8,25,26</sup>.

It is now recognised that this binary classification of phenotypes as Mendelian or complex is an oversimplification and that a continuum spanning from monogenic to polygenic disorders exists<sup>17</sup>. Traits that bridge the two are often referred to as oligogenic. This additional classification stemmed from the fact that not all individuals with known Mendelian mutations presented with the expected phenotype, termed incomplete penetrance, suggesting the presence of a few modifier genes<sup>17,27,28</sup>. Examples include phenylketonuria, cystic fibrosis and Hirschsprung disease, traits that were classically considered to be Mendelian<sup>17,27,28</sup>.

### 1.1.2 Genetic linkage studies

Early success in identifying genetic factors responsible for human trait and diseases, was predominantly for monogenic disorders through the use of genetic linkage studies, first proposed in 1980<sup>18,29</sup>. As of 2003, about 1,200 genes were linked to Mendelian traits<sup>18</sup>. Genetic linkage relies on a definitive phenotype repeatedly occurring within a family that does not affect all members (segregation) and that during meiosis, physically close genes remain in linkage whilst those that are further apart are less likely to cosegregate due to recombination events<sup>18,30,31</sup>. This enables, with the use of genetic markers (approximately spaced uniformly along the genome), identification of genomic regions that are co-inherited in diseased individuals more often than is expected by chance that are not co-inherited in non-diseased individuals<sup>32</sup>. The allelic configuration of a set of genetic markers along a single chromosome (or a section of a chromosome) is defined as a haplotype (Fig. 1.2b) and as result diseased individuals will share the same haplotype at the co-inherited genetic markers<sup>25</sup>.

Typically, these analyses identify large causal genomic regions, which can be refined in subsequent follow up studies using a denser panel of genetic markers in the

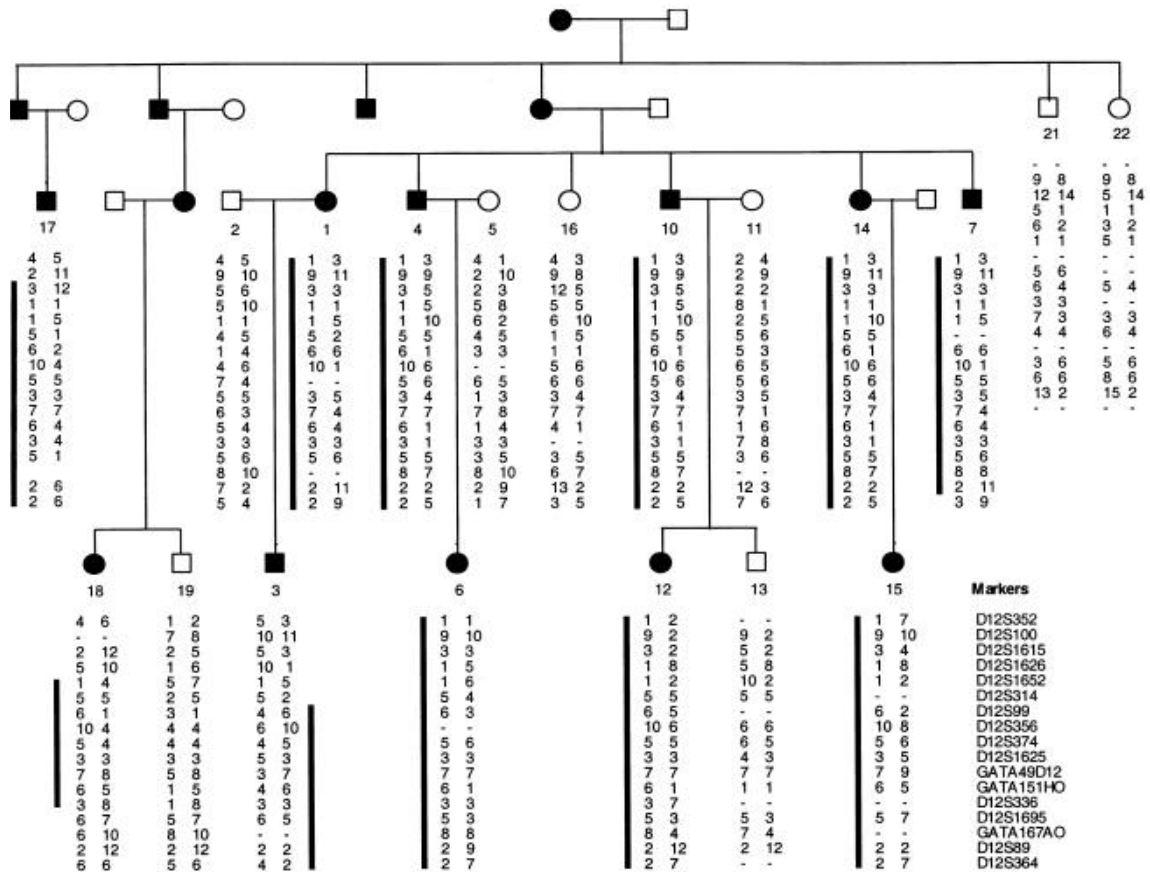
identified region<sup>18</sup>. Nevertheless, the resolution of such analyses ranges between 1 and 10 cM corresponding to between a few and a few hundred candidate genes<sup>18</sup>.

One such example that illustrates this approach, was the identification of a 7 cM region on chromosome 12 for Ehlers-Danlos syndrome type VIII (EDS-VIII), using a large Swedish pedigree<sup>31,33</sup>. In this study, 400 hundred microsatellite markers spaced approximately every 10 cM along the genome were analysed in 11 affected individuals, identifying a putative causal region on chromosome 12. This region was further examined using a denser panel of markers (17 markers spanning a 30 cM region) and additional family members, including eight unaffected individuals to confirm and refine the region<sup>33</sup> (Fig. 1.1). Meiotic recombination in individuals three and 18 results in the identification of a 7 cM region flanked by markers ‘D12S314’ and ‘D12S1695’ as likely causal (Fig. 1.1).

Whilst there was some success in using linkage studies to identify genetic regions involved in complex traits and diseases, including *NOD2* for inflammatory bowel disease<sup>34–38</sup> and *BRCA1* for breast and ovarian cancer<sup>38,39</sup>, the relative lack of success implied that the genetic mechanisms responsible for complex diseases were different to those responsible for rare disorders<sup>26</sup>. That is, complex disorders are not driven by rare variants with large effects. This difference in architecture combined with the use of very small sample sizes, inherently restricted by the size of the pedigree analysed, meant that genetic linkage studies were underpowered to detect regions with significant linkage for complex traits and diseases<sup>38</sup>. This led to both the use of suggestive rather than significant thresholds, resulting in many reported false positive findings and studies that focussed on detecting regions that were not associated with the trait under study, sometimes referred to as exclusion mapping<sup>32,38,40–42</sup>.

### 1.1.3 Common disease/common variant hypothesis

The observation that complex disorders are not driven by rare variants with large effects, led to the common disease/common variant (CD/CV) hypothesis<sup>26,43</sup>, which postulated that common traits are driven by genetic variation that is common in the population. If true, based on the population prevalence of complex traits, the effects of these variants would be small relative to the effects observed for rare variants<sup>26</sup>. Consequently, to explain the estimated heritability (estimated fraction of phenotypic variation explained by genetics<sup>44,45</sup>; see Section 4.1 for further details) of such common traits, many genetic variants must influence complex disease



**Fig. 1.1 Genetic linkage analysis to identify the genetic region linked to EDS-VIII in a Swedish pedigree** | Swedish pedigree of individuals, with those affected by EDS-VIII represented by black symbols and those unaffected by white symbols (square = male, circle = female). Haplotypes for 17 markers on a region of chromosome 12, with black bars next to the marker alleles indicating that the disease segregating haplotype is present in affected individuals. Meiotic recombination in individuals three and 18, results in the identification of a putative causal region, flanked by markers ‘D12S314’ and ‘D12S1695’. Figure as shown in Rahman *et al.*<sup>33</sup>.

susceptibility<sup>26</sup>.

This hypothesis was described as early as 1996, when it was suggested that family based linkage studies would be underpowered to detect variants with more modest effects<sup>46</sup>. Instead it was proposed that using linkage disequilibrium (LD), the observed departure from the expected co-occurrence of two genetic markers based on their observed frequencies (i.e. non-zero correlation between the two genetic markers), combined with larger population based studies would be more suitable<sup>46,47</sup>. This latter design was termed a genomic association study (now referred to as genome-wide association study)<sup>46</sup>.

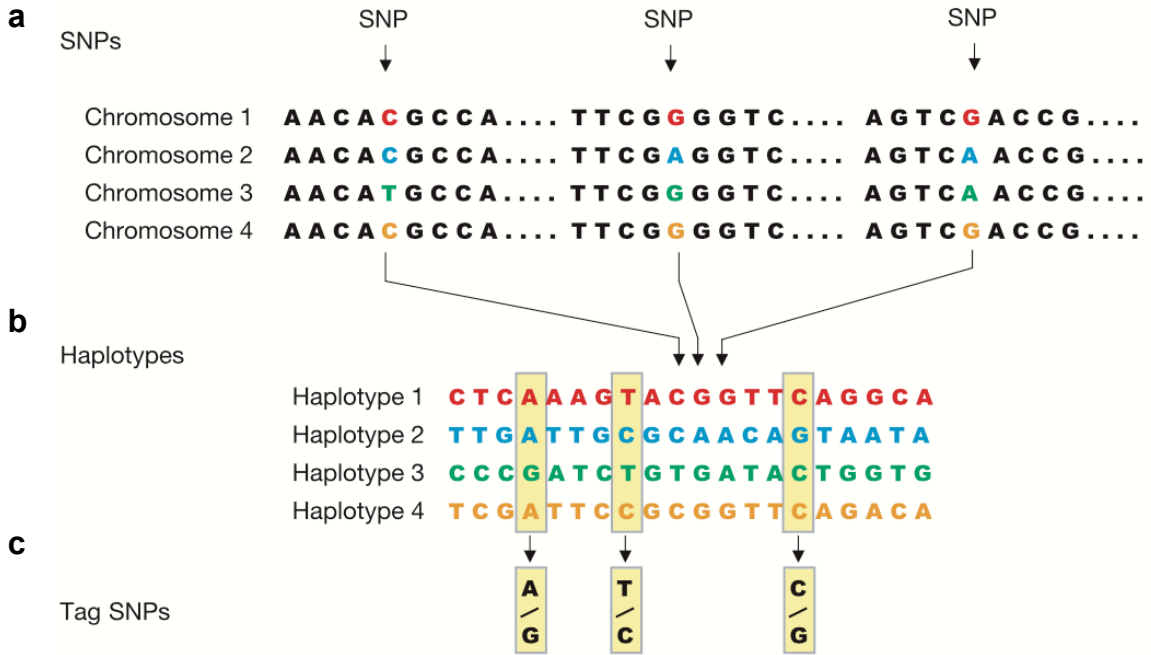
These population based association studies can be viewed as a special case of family based linkage studies, in which the population studied can be viewed as an extended pedigree (due to common ancestors). A greater number of meiotic recombinations will have occurred between the analysed distant relatives, such that LD regions are much smaller than within pedigrees of close relatives, thus requiring a denser panel of genetic markers to be examined<sup>48</sup>.

When proposed, implementation of genomic association studies was not possible. This was due to the lack of information surrounding the location and density of the genetic polymorphisms and the lack of knowledge of the LD between genetic variants across different populations, required for such analyses. In addition, the technology required to genotype thousands to millions of markers in a single experiment for the larger required sample sizes was not available<sup>8,46,49</sup>.

#### **1.1.4 Characterising the LD structure of the human genome**

The International HapMap Project was set up with the goal of characterising the LD structure of the human genome, such that genomic association studies would be feasible. The project initially focussed on 270 samples from four different populations<sup>50</sup> at approximately 1.6 million single nucleotide polymorphisms (SNPs; defined as a DNA sequence variation at a single nucleotide<sup>25,26</sup>, Fig. 1.2a) with minor allele frequency (MAF)  $> 5\%$ <sup>8</sup> and was later expanded to consider 11 populations at 3.1 million SNPs<sup>8,26,51</sup>. The generated data enabled calculation of the LD (Pearson's correlation coefficient squared,  $r^2$ , is commonly used) between SNPs within the genome, effectively describing the chance that two SNPs will be inherited together<sup>26</sup>. Subsequently, this enabled the identification of haplotypes (Fig. 1.2b) and thus a minimal set of SNPs that capture the majority of the haplotype diversity (common variation) within a population, known as tag SNPs (Fig. 1.2c)<sup>50</sup>. Tag SNPs are

population specific, with approximately 500,000 variants in Europeans and up to one million variants in non-Europeans required to capture  $> 80\%$  common variation ( $MAF > 5\%$ ) with LD,  $r^2 > 0.8$ <sup>26,49</sup>.



**Fig. 1.2 Identification of tag SNPs** | (a) Four versions of the same chromosome region, spanning 6,000 bases of DNA, containing three SNPs (i.e. DNA sequence variation across individuals at three single nucleotides). (b) This combination of SNPs defines four different haplotypes, defined as a particular combination of alleles at nearby SNPs. (c) Genotyping just three SNPs across the 6,000 bp region of DNA is sufficient to identify these four haplotypes uniquely and thus these SNPs are defined as tag SNPs. Figure as shown by The International HapMap Consortium, 2003<sup>50</sup>.

These tag variants are used to conduct indirect association studies, where associations are tested for only at the tag variants. As a result, any identified significant associations can be due either to the tested tag SNP or caused by a variant that is in high LD with the tested tag variant. Indirect association studies are much cheaper than conducting direct association studies, where association at each common SNP is tested for<sup>26,48</sup>. Using such tag SNPs, reduces the genetic region associated with a trait to approximately 10 – 100 kb (which is often just a few genes) from the 5 – 10 Mb (which can contain tens to hundreds of genes) regions that were identified with genetic linkage studies<sup>8</sup>.



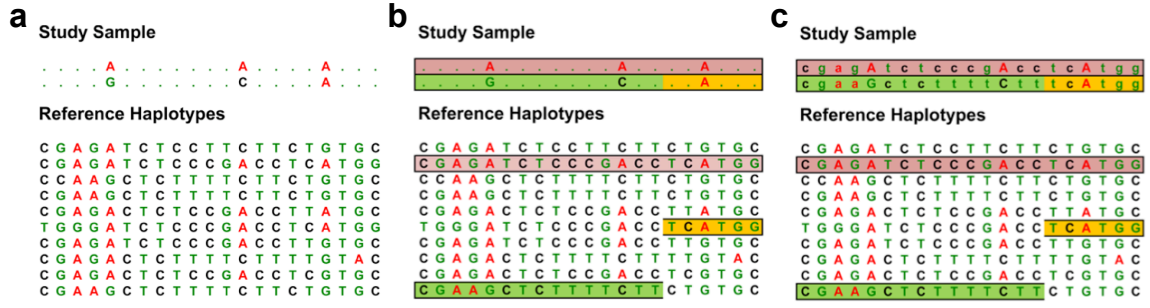
### 1.1.5 Genome-wide association analyses

The data generated by the International HapMap Project combined with development of appropriate chip-based microarray technology, enabling simultaneous genotyping of more than one million SNPs, led to the first wave of genome-wide association studies (GWAS)<sup>26</sup>. Briefly, GWAS are a hypothesis free approach to test for significant correlation between genetic variants (one by one across the entire genome) and a trait of interest. The first successful GWAS conducted in 2005 for age-related macular degeneration using 96 cases and 50 healthy controls, tested for associations at  $\sim 100,000$  SNPs<sup>52</sup>. This was promptly followed by GWAS for Crohn's Disease<sup>53</sup>, myocardial infarction<sup>54</sup>, inflammatory bowel disease<sup>55</sup> and type 2 diabetes<sup>56</sup>. A landmark well-designed study was conducted in 2007 by the Wellcome Trust Case Control Consortium (WTCCC), which demonstrated that a common set of controls for all diseases could be used instead of matching cases and controls for each trait, making the study very cost effective; explicitly, associations using a shared set  $\sim 3,000$  common controls and  $\sim 2,000$  cases (at the time regarded as a large sample size) for each of seven traits were tested for, using a panel of SNPs with good coverage across the genome<sup>8,49,57</sup>.

Whilst these early GWAS were based on the directly genotyped tag SNPs (described in Section 1.1.4), from 2008 onwards, utilisation of methods to fill in missing genotype data became increasingly common practice, referred to as imputation<sup>58</sup>. Imputation is a statistical technique to infer genotypes. The general idea is to compare genotyped samples to a reference panel of individuals of similar ancestry (Fig. 1.3a) to identify short haplotype stretches that are shared between the genotyped and reference samples (Fig. 1.3b). This then allows missing genotypes to be filled in (Fig. 1.3c)<sup>58</sup>. Initially, data from the HapMap Project was used as a reference panel, with for example the CEU panel used to impute samples of European ancestry<sup>58,59</sup>.

Imputation is now also used to explore association effects of rare variants (see Section 1.2.1 for further details of imputation of rare variants, including examples of current reference panels used).

Genotypes may be missing either due to poor quality calls or because they were not directly typed due to the design of the microarray chip. Consequently, imputation allows pooling of results across multiple studies that use different microarray chips (both in design and content) without the need to restrict attention to the overlapping set of SNPs that are examined in all studies, known as meta-analyses (I refer the reader to Evangelou *et al.* for a review on such methods<sup>60</sup>). Meta-analyses result in larger sample sizes and thus in theory greater power to detect associations.



**Fig. 1.3 Imputation of genotype data** | (a) Comparison of samples genotyped at a small number of genetic variants within a given region and a reference panel of individuals that have been densely genotyped in the same region, to (b) identify regions that are shared between the genotyped study samples and individuals on the reference panel, such that (c) unobserved genotypes in the study samples can be filled in based on observed genotypes in the reference panel. Figure adapted from Li *et al.*<sup>58</sup>.

Furthermore, imputation combines evidence across multiple directly genotyped SNPs, such that imputed SNPs may better tag the true causal variant than any of the individually genotyped SNPs. Therefore, imputation can lead to the identification of additional association signals that are not detectable using the raw genotyped data<sup>58</sup>. For example, Kathiresan *et al.* identified a common variant (*rs6511720*) in the *LDLR* gene that is strongly associated with low density lipoprotein cholesterol levels. This was not identified in an initial analysis using directly genotyped variants<sup>58,61</sup>; this is because the best proxy SNP (*rs12052058*) included on the Affymetrix array used for genotyping was in low LD,  $r^2 = 0.21$ , with *rs6511720*<sup>58</sup>.

Initially, GWAS focussed on complex phenotypes with binary outcomes, using a case control design. Quantitative traits have since become increasingly popular to use as phenotypes. This is partly due to the fact that the onset of many diseases is time dependent, meaning that some individuals selected as controls may later become cases and hence the control group actually contains a mixture of cases and controls, resulting in some loss of power<sup>26</sup>. In addition, some binary traits are defined by thresholding a continuous variable, with the threshold somewhat arbitrary such that an individual with a value marginally greater than the cutoff is said to be ‘diseased’ whilst an individual just below the threshold is defined as ‘healthy’. Examples include defining obesity based on a BMI threshold and diabetes based on fasting glucose levels. This thresholding approach results in loss of information regarding phenotypic similarity and as a result, using the binary outcome is likely to be less powered than using the underlying quantitative trait. Moreover, quantitative traits are often closer to the underlying biology, providing greater insight into the mechanisms underpinning trait or disease development and the results from using

continuous traits may be more directly interpretable, in that it is possible to quantify the average difference in trait outcome per risk allele carried<sup>26</sup>.

Traditionally, the effect of an allele was either classified as dominant, when the presence of one allele masks the effect of the second allele or recessive, in which case two copies of an allele are required for an effect. For both binary and quantitative trait study designs, there are multiple ways in which the genotype data can be encoded, allowing for different assumptions regarding the variant impact on a phenotype. This includes the ability to model dominant or recessive effects or alternatively, that each additional copy of a non-reference allele increases (decreases) disease risk either multiplicatively or additively<sup>26,48</sup>. The allelic additive encoding model is often used in practice since it has good power to detect both additive and dominant effects but it can be underpowered to detect recessive effects<sup>26,62</sup>.

Regression based models are often employed for association testing, with case control designs based on logistic regression whilst quantitative traits are based on linear mixed models<sup>26</sup>. These models can be cast as:

$$\text{logit}(\mathbf{y}_D) = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_G + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (1.1)$$

where  $\mathbf{y}_D$  is an  $N \times 1$  binary phenotype vector, capturing for example disease status or:

$$\mathbf{y} = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_G + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (1.2)$$

where  $\mathbf{y}$  is an  $N \times 1$  quantitative phenotype vector. In both cases,  $N$  is the number of samples used in the analysis,  $\mathbf{W}$  is an  $N \times a$  covariate design matrix and  $\boldsymbol{\alpha}$  is an  $a \times 1$  vector of corresponding effects,  $\mathbf{x}$  is an  $N \times 1$  genotype vector for the focal variant (which depending on the encoding can be used to test different assumptions of the variant effect as described above),  $\beta_G$  is the marginal genetic effect and  $\boldsymbol{\epsilon}$  is an  $N \times 1$  noise vector, modelled as random effect following the multivariate normal distribution:

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}_N), \quad (1.3)$$

where  $\mathbf{I}_N$  is the  $N \times N$  identity matrix.

An association test, tests whether  $\beta_G \neq 0$  and different statistical tests, for example the likelihood-ratio test (LRT), score test or wald test, can be employed to achieve this (I refer the reader to Rao *et al.*<sup>63</sup> and Xing *et al.*<sup>64</sup> for further details on these

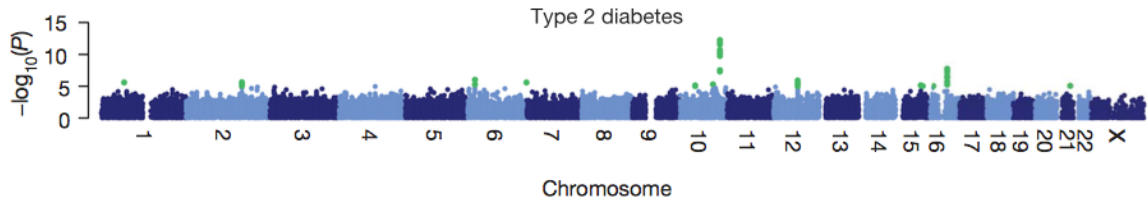
different statistical tests). Formally the association hypothesis test is:

$$H_0 : \beta_G = 0 \quad (1.4)$$

vs

$$H_1 : \beta_G \neq 0. \quad (1.5)$$

GWAS results are often visualised using Manhattan plots, with the negative log P value plotted on the y-axis against the corresponding genomic position ordered by chromosome and position on the x-axis (Fig. 1.4). Peaks on these plots represent loci (multiple variants in LD) that display evidence of association with the analysed phenotype. Variants are deemed to be significantly associated with a trait if they exceed an appropriately chosen P value threshold. Due to LD structure, if a peak arises from a single variant, this is usually indicative of a false positive finding.



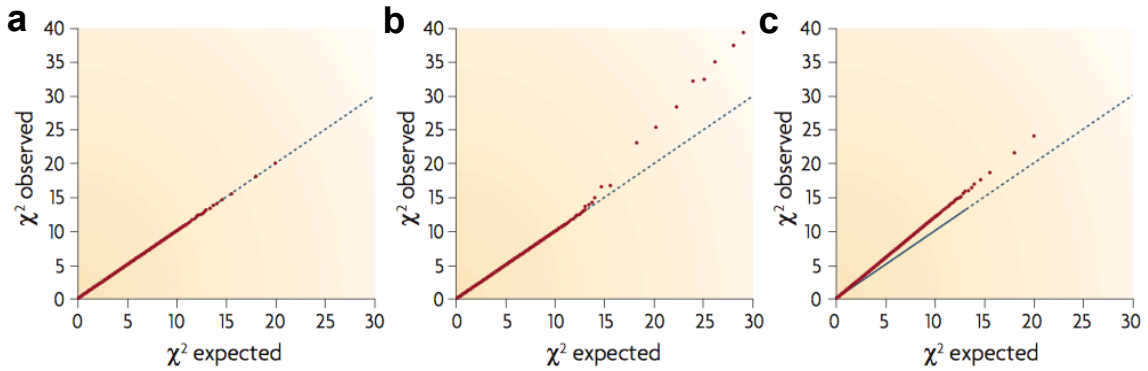
**Fig. 1.4 Example Manhattan plot** | Manhattan plot for type 2 diabetes from the Wellcome Trust Case Control Consortium Study, displaying the negative log P value (y-axis) for each tested variant against the genomic position ordered by chromosome and position (x-axis). Alternate chromosomes are represented by different shades of blue and SNPs displaying evidence of association (in this example  $P < 1 \times 10^{-5}$ ) are highlighted in green. Adapted from the Wellcome Trust Case Control Consortium, 2007<sup>57</sup>.

## 1.1.6 Caveats of association analyses

### Population structure

It was recognised, even before the first GWAS was conducted, that there was a possibility of identifying false positives or that true positives may be masked when using population based association studies instead of family based linkage studies, due to confounding effects<sup>25,48</sup>. In particular, this is because both phenotypic prevalence and allele frequencies vary across different populations, which may result in the identification of variants that are indirectly associated with the phenotype of interest due to ethnicity or population substructure<sup>25,26,48,65</sup>.

Detection of confounding is often assessed by examining the departure of test statistics from the expected null distribution. This can be visualised through the use of quantile-quantile (QQ) plots, with the observed test statistics (or corresponding negative log P values) plotted on the y-axis and the expected test statistics (or corresponding negative log P values) plotted on the x-axis<sup>8</sup> (Fig. 1.5). Some departure from the null distribution in the extreme tail of the test statistics (or corresponding negative log P values) is expected when true association signals are present (Fig. 1.5b), whilst departure from the null due to confounding effects is expected to affect the test statistics (or corresponding negative log P values) globally across the genome, resulting in earlier departure from the null distribution (Fig. 1.5c). A commonly employed measure to assess the presence of confounding is genomic control,  $\lambda_{GC}$ , where the median test statistic is compared to the expected median. If  $\lambda_{GC} \approx 1$ , it is deemed that confounding is not present, whilst if  $\lambda_{GC} \gg 1$ , the presence of confounding is likely, resulting in a high number of false positive findings<sup>66</sup>. It should be noted that larger inflation factors may be observed in GWAS of large sample size for complex traits such as height, due to substantial polygenic signal<sup>67,68</sup>.



**Fig. 1.5 Example quantile-quantile (QQ) plots** | Examples of quantile-quantile (QQ) plots, displaying the expected quantiles (x-axis) versus the observed quantiles (y-axis) of the test statistics under the null (blue line) and observed (red circles) in (a) when no associations are present, such that there is no departure from the null distribution, (b) in the presence of genetic association, such that there is departure from the null in the tail of the test statistics and (c) in the presence of confounding factors, such that there is constant departure from the null distribution. Adapted from McCarthy *et al.*<sup>69</sup>.

Various solutions have been proposed to correct for observed confounding effects. An early solution was genomic control, where the test statistics at all tested variants are adjusted based on the calculated  $\lambda_{GC}$ <sup>66</sup>. However, this method works on the assumption that the tested variants are uniformly affected by the confounding factors, which in the majority of cases will be a gross simplification<sup>8,66,70</sup>.

An alternative method that attempts to correct the underlying problem is STRUCTURE, which assigns individuals to discrete populations subgroups and then combines the evidence for association across the different subgroups<sup>71</sup>. However, this method does not scale with sample size and is also highly sensitive to the number of defined population clusters<sup>70</sup>.

This led to the development of a solution, that is widely used today, which does not require defining discrete subpopulations but rather assumes a continuum. This is achieved through the use of principal component analysis (PCA) of the genetic data, to infer axes of variation due to population structure. The identified top principal components (PCs) are either first regressed on the phenotype, with subsequent association testing performed using the phenotypic residuals or the PCs are included as covariates (i. e. in the design matrix  $\mathbf{W}$  in Eqs. 1.1-1.2) in the linear regression model used for association analyses<sup>70</sup>. However, the top PCs only capture the axes that explain the greatest amounts of variation and hence do not capture more subtle relationships such as that between closely related individuals. Therefore, when PCs are used to correct for population structure, closely related individuals should be removed from the association analyses, prior to PCA calculations.

An alternative, more computationally demanding strategy, that can account for both population structure due to closely related individuals and the presence of subpopulations, is the use of linear mixed models (LMMs)<sup>72,73</sup>. Instead of calculating principal components, the genetic data is used to estimate an  $N \times N$  kinship matrix,  $\mathbf{R}$ , that describes the genetic similarity between pairs of individuals (see Hayes *et al.*<sup>74</sup> for a description of a commonly used similarity measure used to generate  $\mathbf{R}$ ). This genetic similarity is modelled through the use of an additional random effect term in the linear model, described by Eq. 1.2 (and equivalently, for logistic regression described in Eq. 1.1), as follows:

$$\mathbf{y} = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_{\text{G}}}_{\text{G}} + \underbrace{\mathbf{u}}_{\text{Confounding}} + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (1.6)$$

where  $\mathbf{u}$  is modelled as a random effect, following the distribution:

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \mathbf{R}). \quad (1.7)$$

It can be shown that the LMM approach is theoretically equivalent to the PC approach when all PCs are regressed or included as covariates (see Hoffman *et al.*<sup>75</sup> for details), explaining why LMMs are able to account for more subtle population structure than the PC approach. However, regressing all PCs or including all PCs

as covariates is not feasible in practice.

The advantages of LMMs for association analyses, has resulted in research focussed on improving the efficiency of these methods. Whilst naively the LMM scales cubically with the number of samples, there are now methods available that scale linearly with the number of samples, once some up front computations that scale quadratically with the number of samples have been performed. These include FaST-LMM<sup>76</sup>, BOLT-LMM<sup>77</sup> and LIMIX<sup>78</sup>, the latter built using a flexible framework such that different types of testing procedure can be easily and efficiently implemented, e. g. multi-trait set tests<sup>79</sup> and interaction set tests<sup>80</sup> (see Section 1.2.3).

### Testing multiple variants

Given that a much greater number of variants are tested for associations in population based studies than family-based linkage analyses, a second caveat of population based association testing, is the need to use appropriate multiple testing adjustments such that the number of reported false positive findings is controlled<sup>8,26</sup>. Adjustment for testing multiple hypotheses is required since some tests statistics will be very extreme by chance when many hypotheses are tested<sup>81</sup>.

One way to account for multiple testing is to control the family wise error rate (FWER) at a given level  $\alpha$  (often  $\alpha = 0.05$ ), such that the probability of making one or more false discoveries across all tested variants is less than  $\alpha$ <sup>81</sup>. The Bonferroni correction is commonly used, such that the  $i^{\text{th}}$  variant is significantly associated with a phenotype, only if  $P_i \times m < \alpha$ , where  $m$  is the number of tests conducted ( $P_i \times m$  is often referred to as the Bonferroni adjusted P value)<sup>26</sup>. This correction method assumes that the test at each variant is independent of the tests at all other variants. Hence, in the presence of LD, this multiple testing correction can be conservative. Methods have been developed to calculate the ‘effective’ number of independent tests, accounting for LD, such as eigenMT<sup>82</sup>. The ‘effective’ number of tests is given by the number eigenvalues (from PCA) required to explain a certain percentage of the total variation (a suggested value is 99%)<sup>82</sup>. The widely accepted genome-wide significance threshold of  $5 \times 10^{-8}$  was derived using this Bonferroni correction, after calculating the number of independent tests using the SNP data based on individuals of European ancestry, generated by the International HapMap Project<sup>81,83,84</sup>. Since the International HapMap Project considered only common variants ( $\text{MAF} > 5\%$ ), this threshold is unlikely to be appropriate if variants with lower MAFs are included in the association analysis, as is now common practice or alternatively, if different population ancestries are considered due to differences in

LD structure. For example, as described in Section 1.1.4, a much greater number of tag SNPs are required for Africans than non-Africans, indicative of a greater number of independent tests.

However, the FWER can be very conservative when thousands or millions of tests are conducted<sup>85</sup>. Therefore, an alternative method to account for multiple testing is to control the expected false discovery rate (FDR) at a given level  $\alpha$  (often  $\alpha = 0.05$ ) and this less stringent procedure controls the expected number of ‘discoveries’ (significant results) that are false positives<sup>81,85</sup>. This approach was formally described in 1995, with the algorithm provided known as the Benjamini-Hochberg FDR<sup>86</sup>. Various closely linked extensions and alternatives have been since proposed including the local FDR<sup>87</sup>. This local FDR is a slightly less stringent formulation of the widely used Benjamini-Hochberg adjustment procedure (with equivalence under the prior assumption that the fraction of null hypotheses is one). With this approach, the P value of each variant is adjusted based on its rank (with increasing P values corresponding to greater rank), such that a variant is significantly associated with a phenotype if  $\frac{P_i \times m}{r_i} < \alpha$ , where  $m$  is the number of tests conducted and  $r_i$  is the rank of the  $i^{\text{th}}$  variant ( $\frac{P_i \times m}{r_i}$  is often referred to as the Benjamini-Hochberg adjusted P value)<sup>87</sup>.

## Replication of significant results

Where possible, a further strategy commonly employed in an attempt to reduce the number of reported false positive results, is replication of association findings using an independent data set<sup>8,26</sup>. Ideally, this independent data set should be obtained from the same population as that used for the original study; this is because it is not necessarily expected that the identified associations will be observed in other populations due to allele frequency differences<sup>26</sup>. However, other populations can be used to asked questions about the population specificity of the identified findings<sup>26</sup>. Replication studies usually focus on the subset of interesting variants that were identified in the primary analysis, testing for associations only at these variants or proxy SNPs in high LD, if the original SNP is not available in the replication study<sup>8,26</sup>. Since a smaller number of variants are considered in the replication analysis, the P values are not required to be as stringent as for the primary analysis but the variant effect should have a consistent direction of effect with the discovery analysis<sup>8</sup>.



## 1.2 Advances in association analyses

Following initial success in identifying loci associated with complex trait and disease risk, it was recognised that the common variants identified explained only a small fraction of the expected trait heritability, typically  $< 5\%$  as of 2009<sup>59,88–92</sup>. As a result, to further increase knowledge of complex traits and diseases, various improvements outlined in greater detail below have been implemented. This includes methodological advances and an increase in the scope of data collection. Data breadth, has increased in terms of the the density of genetic variants, the number of individuals and the number of phenotypes analysed; the latter includes many phenotypes that can be viewed as mediating traits such as gene expression data and quantitative traits that are predictive of disease outcome.

### 1.2.1 Rare variants

Given the substantial unexplained heritability considering only common variants ( $MAF > 5\%$ ), attention turned to the role of low frequency ( $1\% < MAF < 5\%$ ) and rare variants ( $MAF < 1\%$ ) in common disease<sup>89</sup>. Improvements in available technology, in particular next generation sequencing (NGS)<sup>92,93</sup> paved the way for exploring the role of such variants in common traits and diseases. In particular, the 1000 Genomes Project employed this technology to characterise rare variants using low coverage whole genome sequencing (WGS) and deep whole exome sequencing (WES), initially examining 1,092 individuals from 14 ancestries and later expanding to 2,504 genomes from 26 populations<sup>94,95</sup>.

Genotypes of less common variants for association testing can either be obtained directly through WGS or WES or alternatively, due to the relatively high costs of these direct sequencing methods, can be obtained through imputation (see Section 1.1.5 for details). Since, imputation accuracy decreases with MAF but increases with the size and the use of an ethnically matched reference panel, choosing an appropriate reference panel is particularly important for rare variant imputation<sup>89,92,96,97</sup>. Examples of population specific WGS reference panels include UK10K<sup>98</sup>, deCODE<sup>99</sup>, Genome of the Netherlands<sup>100</sup>, HELIC-MANOLIS<sup>101</sup> and SardiNIA<sup>59,102,103</sup>. There are also ongoing efforts to combine all available WGS data to generate a single large reference panel, known as the Haplotype Reference Consortium (HRC) panel<sup>59,104</sup> and also to generate a reference panel for sub-Saharan Africa, as part of The African Genome Variation Project<sup>105</sup>.

The analysis of rare variants resulted in the need for new methods to increase the statistical power to identify associations between these variants and phenotypes<sup>96</sup>. These developed methods test for associations between a set of multiple rare variants instead of testing for associations at each variant individually (see Eqs. 1.1-1.2). This aggregation across multiple rare variants can increase the observed frequency such that there is sufficient power to detect an association and in addition reduces the multiple testing burden (see Section 1.1.6)<sup>106</sup>. These methods are often referred to as set tests and can be categorised as burden, adaptive burden, variance component or combined burden and variance component tests<sup>96</sup>.

Briefly, burden tests<sup>107–111</sup> collapse information across multiple genetic variants, resulting in a single combined genetic score and then test for association between this combined genetic score and the phenotype using standard regression methods (see Eqs. 1.1-1.2). As a result, these tests work under the assumption that all variants contribute to the observed phenotype and act with the same direction and magnitude of effect<sup>59,96</sup>. However, many rare variants will not confer an effect on complex trait or disease risk. This led to the development of adaptive burden tests which relax this assumption by incorporating information about individual variant effects<sup>59,96,112</sup>. This can include, switching the reference and alternative allele if there is evidence to suggest that variants have opposite directions of effect and excluding variants that are unlikely to have an effect based on marginal genetic association results (i.e. using a standard linear regression approach as described in Eqs. 1.1-1.2). In addition, variants included within the set can be weighted based on the SNP effects obtained from these single variant analyses<sup>112–117</sup>. As a result many of these adaptive burden tests rely on a two stage procedure, where in the first stage marginal association tests are conducted and in the second a set test is conducted, which can be time consuming<sup>96</sup>. This is further compounded by the fact that these tests often rely on permutations to obtain P values, which is computationally intensive. Furthermore, these adaptive burden tests can also suffer from unstable estimates of individual rare SNP effects<sup>96</sup>.

An alternative that does not constrain dependence in the magnitude and direction of effect between variants within a set are variance component tests, which use a random effects framework<sup>59,96,118–120</sup>. The Sequence Kernel Association Test (SKAT)<sup>118</sup> is a commonly employed rare variant method that falls within this category.

Since variance component tests can be more powerful than burden tests if multiple rare variants influence a phenotype with different magnitudes and direction of effect

but burden tests can be more powerful when a number of rare variants within the set act with similar magnitude and with the same direction of effect, tests that combine burden and variance component tests have also been proposed<sup>59,96,121–123</sup>. The Optimal Sequence Kernel Association Test (SKAT-O)<sup>123</sup> falls within this category and is commonly used for rare variant analyses.

As already alluded to, a major associated challenge of set test based approaches is appropriate selection of the variants to include within a set; the majority of rare variants will not be associated with a trait and inclusion of such null variants, can hamper the power of set test approaches. For example, weaker signals were seen when testing for associations between a set of rare variants in genes *BCAM* and *CD300LG* with blood lipids using both burden and SKAT tests than when individual variants were considered<sup>96,124</sup>. However, set test based approaches can lead to substantial power gains, for example a gene based test using SKAT-O for variants in *PLD3* resulted in an association P value of  $1.4 \times 10^{-11}$  with Alzheimers, whilst  $P < 1 \times 10^{-6}$  was not observed for any single variant<sup>96,125</sup>. Commonly, regional or gene-based and/or functional annotations are used to select variants. There exist a number of computational prediction tools to annotate variants within coding regions, including but not limited to SIFT, PolyPhen2 and CADD, although there are often differences in the predicted pathogenicity of variants between the tools, rendering the selection of variants a difficult problem<sup>126–128</sup>. This is an even greater issue when considering rare variants in the non-coding regions, where there exists a greater pool of variants to start with and less functional knowledge available.

### 1.2.2 Increased sample sizes

Observations that heritability estimates were larger when including all common variants instead of only those that were genome-wide statistically significant, implied that common variants with smaller effect sizes not yet identified through GWAS, influence complex traits and diseases<sup>89,90,129</sup>. As sample size is a critical factor that can increase the power to detect variants with smaller effect sizes<sup>89,90,129</sup>, the sample size for GWAS has been increasing. This has been achieved both through the generation of larger cohorts or as mentioned in Section 1.1.5, by combining data across multiple cohorts, often based on summary statistics using appropriate meta-analyses methods (see Evangelou *et al.* for a review on such methods<sup>60</sup>). In addition, if appropriate technology is used to capture rare variants, larger sample sizes simply increase the probability of observing rare variants (i.e. to capture a single individual that carries a variant with population frequency of 0.05% requires

on average a sample size ten times as large as that required to capture an individual that carries a variant with population frequency of 0.5%, assuming perfect detection of rare variants) and thus increase the expected number of individuals within a cohort that carry a rare variant at a specific chromosomal position, hence increasing the subsequent power to detect rare variant associations (see Section 1.2.1)<sup>96</sup>.

Cohorts comprised of the order of hundreds of thousands of participants are now being directly generated. These are largely population based cohorts, that are collecting a multitude of phenotypic data, including electronic medical records (EMR), biospecimens and imaging data in addition to genotypic data. Such biobanks have been developed in a number of countries worldwide, including Iceland, Sweden, Denmark, Latvia, Estonia, Canada, South Korea, Japan, Singapore, China, Mexico, USA and UK<sup>130,131</sup>.

In the UK, the largest of these is UK Biobank, which is a prospective cohort study that recruited ~500,000 British individuals aged between 40 and 69 years at recruitment, between 2006 and 2010<sup>131</sup>. This study is collating a deep repository of phenotypic and environment information, using questionnaires, links to medical records, physical measurements (including accelerometer data), biological samples and imaging<sup>131</sup>. An advantage of prospective cohort studies is that they collect data prior to disease onset and this study supersedes those previously conducted by combining a large cohort size with collection of deep meta-data<sup>131</sup>. Furthermore, directly generating large biobanks can be better than using meta-analyses that combine smaller cohorts since there is increased homogeneity in the definition of the phenotypes and environmental data collated<sup>131,132</sup>.

### 1.2.3 Multi-trait methods

An increase in power to identify variants with small to modest effect sizes can also be achieved through the joint analysis of multiple correlated phenotypes using multi-trait LMMs, where smaller genetic effects are effectively aggregated across multiple traits whilst controlling for population structure<sup>79,133,134</sup>. As well as testing for associations between a variant and considered traits, the underlying null model can be used to partition phenotypic correlation into shared genetic and environmental components across the multiple traits<sup>133,135</sup>. Thus, such multi-trait methods can also increase our understanding of the shared genetic architecture and mechanisms underpinning complex traits and diseases<sup>133</sup>.

Korte *et al.*<sup>133</sup> were the first to apply such methods on a genome-wide scale, testing

for associations with four blood metabolites, considering pairs of phenotypes. Multi-trait methods included within LIMIX, GEMMA and GAMMA provide computational efficiency improvements, enabling such analyses to be applied to tens of traits and thousands of samples<sup>78,134–136</sup>. More recently mtSet which combines such multi-trait tests with variant set tests was developed, compounding the aforementioned advantages of both variant set tests (see Section 1.2.1) and multi-trait modelling, whilst retaining computational efficiency<sup>79</sup>. Although the initial proposed mtSet only allowed for association testing, an extension to this method, iSet can be used to examine the shared genetic architecture<sup>80</sup>.

As well as analysing distinct correlated phenotypes, all of the aforementioned methods can be applied to multiple phenotypes that measure the same trait under different environmental conditions and thus can be used to identify context dependent variation<sup>80,133</sup>. For example, iSet was used to test for stimulus-specific effects on gene expression of primary monocytes, a white blood cell involved in immune response, where the measured gene expression of a specific gene under four different environmental conditions comprised the four phenotypes included in the multi-trait test<sup>80</sup>. However, these methods are limited to categorical environments<sup>80</sup>.

## 1.2.4 Pleiotropy and PHEWAS

Variants can also be associated with multiple traits that are not correlated, referred to as pleiotropy<sup>137</sup> (although increasingly this term is used when a variant is associated with multiple traits, regardless of their correlation<sup>138,139</sup>). Identification of pleiotropic effects may increase our functional knowledge of complex traits and diseases. For example, pleiotropy may occur when a locus affects a molecular mechanisms that alters risk for multiple traits (termed direct biological pleiotropy) or the locus may affect one trait that is a molecular intermediate for a second trait that I henceforth refer to as the outcome trait (termed mediated pleiotropy)<sup>138–140</sup>. The latter can often be deduced through the use of Mendelian Randomisation, a causal inference method<sup>140,141</sup>. Explicitly, variants that are associated with the mediating trait are used as proxy measures for this trait, such that that the analysis is robust to reverse causation (i. e. when the outcome trait alters levels of the intermediate trait) as genetic variants are not modifiable<sup>142</sup>. In this way, Mendelian Randomisation is analogous to randomised control trials, where individuals are randomly placed into different treatment groups such that there is no confounding between the treatment and outcome<sup>143</sup>.

Examples of pleiotropy include variants in *CTLA4* and *PTPN22* that are associated with multiple autoimmune disorders<sup>144–146</sup> and genetic variants near *CDKN2A* and *CDKN2B* associated with type 2 diabetes and coronary heart disease suggesting a shared mechanism or biological pathway<sup>147–151</sup>. A systematic review in 2011 of variants listed in the GWAS catalog<sup>152</sup>, revealed that almost 5% of SNPs and almost 17% of genes or gene regions are associated with more than one distinct outcome or trait<sup>139,151</sup>.

Whilst historically pleiotropic effects were identified through literature searches or catalogs such as the GWAS catalog<sup>152</sup>, the generation of large biobanks with links to EMR (see Section 1.2.2) or that have collected multiple phenotypes, for example the PAGE study<sup>151,153</sup> have enabled testing for pleiotropic effects within the same cohort of individuals, referred to as phenome-wide association studies (PHEWAS). Rather than focussing on a phenotype of interest as is the case for GWAS, PHEWAS focus on variants of interest and test for associations with all available phenotypes.

The first PHEWAS study conducted in 2010, focussed on five variants with previous disease associations, testing for associations with 733 phenotypes; this study replicated four of seven known variant-disease associations and identified 19 novel associations ( $P < 0.01$ )<sup>154</sup>. This early proof of concept study highlights how PHEWAS can be used to identify possible pleiotropic effects and as a hypothesis generating tool<sup>155,156</sup>. Identification of possible pleiotropic effects provides opportunities for pharamcogenomics, including the use of already available drugs to treat different conditions or as a predictive tool of efficacy or adverse reactions to a drug<sup>151,155–157</sup>. More recently, these types of study have been extended to a genome-wide scale, testing for associations between 635,525 variants and 541 traits and 25 clinical phenotypes<sup>158</sup>.

One challenge of PHEWAS is the associated multiple testing burden (see Section 1.1.6), which increases with the number of traits analysed<sup>138,157–159</sup>. However, the number of tests does not necessarily scale linearly with the number of traits as in many cases there will exist some correlation between the traits included in the analysis; thus, the tests across different traits are not independent of one another. The effective number of independent tests can be estimated analogous to GWAS where correlation between variants exists, using PC based methods, such as eigenMT<sup>82</sup> (see Section 1.1.6). Nevertheless, the significance threshold is likely to be stringent in many cases, as the sample size per trait and for binary traits the number of affected individuals, can be highly variable<sup>157,158</sup>.

## 1.3 What about the other components of complex traits and diseases?

### 1.3.1 Other contributing factors

Whilst the advent of GWAS and the subsequent advances have increased our understanding of the genetic factors underpinning complex traits and diseases, such that as of May 2018, 69,000 significant SNP-trait associations were listed in the GWAS catalog<sup>152</sup>, the majority of these studies assume that phenotypes are the result of additive variants effects.

Environmental factors can also have a direct effect on phenotypic risk; for example, blood cadmium, lead, polychlorinated biphenyls, dietary nutrients and particulate matter have all been associated with increases in blood pressure<sup>160–169</sup>. Linear models, analogous to that described in Eq. 1.2, with the genotype vector ( $\mathbf{x}$ ) replaced by an environmental exposure vector are commonly used to identify such environmental factors. The majority of such studies analyse the impact of a single environmental factor meaning that the results of these studies may be prone to selection biases and false positive reporting<sup>160</sup>. To overcome such issues there has been a recent drive to systematically evaluate the impact of environmental exposures, referred to as environment-wide association studies<sup>160,170,171</sup>, made possible by the in-depth collection of environmental data such as that collated by the National Health and Nutrition Examination Survey in the USA.

As well as acting independently of one another, variants and environmental exposures can be dependent on one another. In some cases, this may mean that the effect of a variant on a phenotype can only be identified once the effect of another variant or an environmental variable has been accounted for<sup>132,172–177</sup>. This dependence is usually referred to as an interaction effect, with dependence between multiple genetic variants termed epistasis, whilst dependence between a variant and an environmental exposure is termed a genotype-environment interaction ( $G \times E$ ) effect. Interaction effects between variants in the *NAT2* gene and smoking on bladder cancer risk is an example of a  $G \times E$  effect<sup>178,179</sup>.

Whilst the concept of interaction effects has been around for more than a century, the advent of GWAS has enabled feasible exploration of such effects on a genome-wide scale<sup>177,180</sup>. In particular, increased interest in these studies was fostered when the initial wave of GWAS were unable to identify variants that explained substantial amounts of the estimated phenotypic variance<sup>181</sup>.

The identification of additional variants involved in complex traits and diseases through epistasis and  $G \times E$  studies, may provide greater clarity as to which biological pathways are functionally involved in trait onset and progression, as well as highlight potentially novel biological pathways; these pathways may be of interest for drug development<sup>132,177,182,183</sup>. As well as the identification of additional variants associated with complex traits and the ability to explain additional phenotypic variance, insights from interaction analyses may be clinically informative, both to identify the environments with greatest impact and to identify particular subgroups of the population that will most benefit from environmental changes<sup>132,177,183,184</sup>.

### 1.3.2 Epistasis and $G \times E$

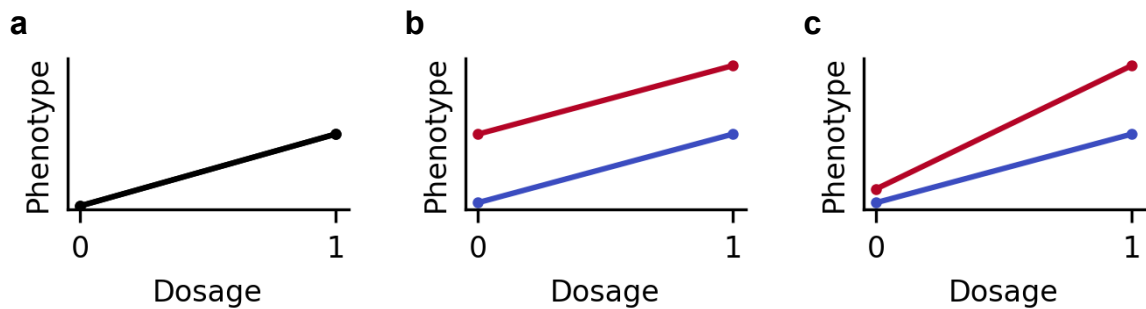
Epistasis was initially described in terms of Mendelian inheritance (see Section 1.1.1) and its presence demonstrated by Bateson in 1909<sup>185</sup>. Specifically it was described as a masking effect, where the effect of one allele, prevents a second allele from exerting its effect (i. e. an extension of the dominance effect)<sup>172,186,187</sup>. In 1918, Fisher defined epistasis as departure of the phenotype from a model that assumes independence of variant effects<sup>2,172,187,188</sup>.

This is the basis of the statistical definition for both epistasis and  $G \times E$  effects that is widely used in the study of complex traits and diseases<sup>182,186</sup>. Explicitly, whilst a variant with a significant genetic effect is determined by a mean phenotypic difference between groups of individuals with different allele dosages (Fig. 1.6a) and a significant environment effect is determined by a mean phenotypic shift that is constant across different allele dosages (Fig. 1.6b), a significant statistical interaction effect is defined when there is a significant difference in the genetic effect between two groups of individuals with different environmental exposures (Fig. 1.6c).

Environment in the context of  $G \times E$  has a broad definition, encompassing microbiota composition, medication, gestational environment, internal mediating factors such as biomarkers, epigenetic features and factors that have a genetically driven component<sup>183,189</sup>.

There are challenges that are common to both epistasis and  $G \times E$  studies. One such challenge is that larger sample sizes are required to detect interaction effects than marginal effects; it has been suggested that sample sizes four times as large are required to detect interaction effects as marginal effects<sup>132,177,182,189,190</sup>. One reason for this is that successful GWAS depends on high LD between the causal variant and the observed tagging variant since the additive variance of the tagging variant will





**Fig. 1.6 Different relationships between allele dosages, environmental exposures and phenotypes** | The relationship between allele dosage and the mean phenotypic value when (a) a significant genetic effect is present, with the variant effect given by the line gradient, (b) in the presence of a direct environment effect where the red and blue lines represent two groups of individuals that are stratified based on environmental exposure levels (either lowly and highly exposed or homozygous reference or homozygous alternative at a second allele that is independently associated with the phenotype) such that the allelic effect of each population subgroup is the same (line gradients are equal) but there is a shift in the mean phenotype that is independent of the focal allele dosage and (c) in the presence of an interaction effect where the red and blue lines represent two subgroups of individuals that are stratified based on environmental exposure levels (either lowly and highly exposed or homozygous reference or homozygous alternative at a second allele) such that the genetic effect in the two groups are different (difference in line gradients).

decrease linearly with LD,  $r^2$ <sup>180</sup>. It follows that the variance explained by pairwise epistasis effects decreases on average with LD  $r^4$ <sup>180</sup>. Similarly, for G×E, the observed environment may be tagging the causal interaction environment, which similarly reduces the power to detect interaction effects. Whilst not directly alleviating this problem, the generation of cohorts with increasing sample size, in particular large scale cohorts such as UK Biobank (Section 1.2.2) should enable the detection of more modest interaction effects.

However, since these large cohorts are also starting to collect multiple environmental variables, if interaction effects with different environmental factors are tested for, there will be an increased multiple testing burden (see Section 1.1.6) for G×E studies that may counteract the gain in power obtained from using larger sample sizes. This multiple testing burden is an even greater problem for epistasis studies<sup>177,186</sup>. Whilst there will be  $\frac{V(V-1)}{2}$  tests, where  $V$  is the number of genotyped (and imputed) variants, when considering all possible pairwise interactions, there is a much greater number of tests when considering higher order interactions<sup>189</sup>. The multiple testing burden associated with exhaustive testing of higher order epistasis effects is likely to outweigh the benefit of considering more complex models<sup>180</sup>. In comparison, typical G×E scans that consider a single environmental variable conduct  $V$  tests, the same number as the corresponding GWAS. Even when multiple environmental variables are considered, typically  $L \ll V$  (where  $L$  is the number of environmental variables) such that the total number of interaction tests ( $VL$ ) is still much smaller than for exhaustive pairwise epistasis scans. As a result, G×E analyses are more tractable than epistasis studies.

Challenges that are specific to G×E tests largely arise due to the fact that measuring environmental exposures is less straightforward than genotyping. In particular, global standards for measuring environmental exposures are generally not defined, such that G×E meta-analyses are likely underpowered and the replication of identified interactions is difficult<sup>182,183</sup>. Even within a cohort environmental measures are more prone to error, in particular when self reporting questionnaires are used or due to changes in behaviour when people know that they are being observed, known as the Hawthorne effect<sup>177,179</sup>. In addition, not knowing which exposures and at what time points (e.g. during in-utero development, childhood development or accumulation over time) these environments matter, with exposures changing over time, makes it difficult to both collect the required data and capture or describe an environmental exposure through the use of single variable generally required for modelling purposes<sup>132,182,183</sup>.

A further difficulty in replicating  $G \times E$  effects arises because environmental exposures can vary substantially with geographical location, meaning that finding an appropriate replication cohort may not be that easy<sup>179,182</sup>. This is compounded by the fact that the identified environmental exposures may be tagging the causal interaction environment and this tagging may vary across different populations (can be viewed similarly to different LD tagging across different populations).

Interpretation of  $G \times E$  is not as straightforward as for epistasis studies. In particular, it is possible that environmental variables can themselves have a genetic component and it might be this genetic component that is driving the observed interaction effect such that the identified  $G \times E$  effect is actually an epistatic effect. If it is the non-genetic component of the environment driving the interaction effect, then the  $G \times E$  effect may be more apparent in subjects of a given age range, which again can make replication difficult<sup>189</sup>. In addition, as greater numbers of environmental variables are tested for interaction effects, it is becoming apparent that the identified environment may not be the driving environment but instead tag an unobserved or compound environmental measure. Given that the environmental search space is not a well defined domain, identification of the causal environment may prove to be very difficult.

Whilst there are some differences in the challenges for identifying epistasis and  $G \times E$  effects, the majority of methods used to detect interaction effects can be applied both to epistasis and to  $G \times E$ , since the statistical definition is the same for both<sup>177</sup>. As  $G \times E$  is the application focus throughout this thesis, for simplicity I will describe commonly used interaction methods in Section 2.1 in terms of  $G \times E$ , noting that for epistasis the environmental variable can simply be replaced by a genetic factor.

Due to the aforementioned shortcomings, there has been limited success in robustly identifying both genotype-environment and epistatic interaction effects. However, some  $G \times E$  effects for human disease risk<sup>191,192</sup> and molecular traits<sup>193,194</sup> have already been identified and there are concerted efforts underway to increase sample sizes and improve the quality of environmental data collected. In particular, this is through the generation of large scale biobanks that collate a multitude of phenotypic and environmental data, such as UK Biobank<sup>131</sup> (see Section 1.2.2), which in theory enables a more comprehensive analysis of genotype-environment interactions.

These large scale datasets with multiple environmental measures available, bring to the forefront a new problem; as a greater number of interaction effects are identified due to the increased power, there is a growing body of evidence to suggest that multiple environments are linked to interaction effects at the same genetic loci for a

given trait. For example, there is evidence of  $G \times E$  effects due to different lifestyle factors, including diet and physical activity at *FTO* on BMI<sup>195–206</sup>.

These environments are often correlated with one another, making it difficult to interpret whether the identified environments are acting independently of one another or tagging the same either observed or an unmeasured environment<sup>179,195</sup>. The latter would invalidate the claims of early studies, which often stated that *the* tested environment is *the* driving interaction environment and that public health policies should be based on these findings<sup>207</sup>. Furthermore, testing for interactions at individual environments makes it difficult to determine which environmental components have the greatest impact on the observed interaction effect<sup>179,195,208–211</sup>.

This has fostered interest to jointly test for interaction effects at multiple environments but established computationally efficient robust methods with available software, prior to this work were not available (see Section 2.2.7 for further details). Many of the advantages of doing so are similar to those of set tests that aggregate over variants (see Section 1.2.1). Namely, set tests that aggregate over environments may increase the ability to detect interaction effects with modest or weak effects at individual environments or alternatively a set of environments may better tag the true driving environment. In addition, if the original analysis plan was to test for interaction effects using different environments (one by one), environment set test based approaches will reduce the multiple testing burden. Furthermore, methods that account for other correlated environments can increase the ability to identify the relative importance of different environmental components for an observed interaction effect.

These benefits, together with the growing availability of large scale biobanks that are collecting deep environmental data, necessitate the need for method development in this direction.

## 1.4 Thesis overview

The overall aim of this thesis is to provide suitable methods to identify multivariate  $G \times E$  and explore these effects across a range of phenotypes using the UK Biobank data.

Specifically, in Chapter 2, I describe the structured linear mixed model (StructLMM), a novel computationally efficient multivariate  $G \times E$  framework that can be used to test for interaction effects. In addition, the same framework, similarly to existing

interaction methods can be used to test for joint association effects, i.e. test for associations whilst accounting for possible heterogeneity in variant effects across individuals due to differences in environmental exposures. I show through the use of simulation experiments that StructLMM is robustly calibrated and in general, better powered than existing interaction and association tests.

In Chapter 3, I present an application of StructLMM, where I identify significant interaction effects with 64 lifestyle-based factors for BMI using the UK Biobank data. In addition, I show that the StructLMM association test can be used to identify loci with genotype-environment contributions. Subsequently, I explore characteristics of significant interaction loci, including the fraction of the genetic variance that is explained by  $G \times E$  and the environmental profiles that increase or decrease phenotypic risk, using methods that are implemented as part of StructLMM.

Finally, in Chapter 4, I apply the StructLMM interaction test to multiple cardiometabolic traits using the UK Biobank data, facilitating exploration of the shared  $G \times E$  architecture. Additionally, I provide preliminary estimates of the amount of phenotypic variation that can be explained by  $G \times E$  effects, relative to marginal association effects.



# Chapter 2

## StructLMM: a linear mixed model approach to study multivariate genotype-environment interactions

### 2.1 Introduction

As outlined in Section 1.4, there is evidence that a single genetic locus may interact with multiple environments to influence the outcome of a given phenotype. For example, it has been shown that *FTO* interacts with physical activity<sup>195–197,212,213</sup>, diet<sup>195,196,206,214</sup> and smoking<sup>196</sup> to impact BMI. Difficulties in interpreting results due to correlation between these environmental factors combined with the increasing availability of large datasets, such as UK Biobank<sup>215</sup>, that collect deep phenotype, environment as well as genotype data, necessitates a need for multi-environment  $G \times E$  methods.

A current way to detect interaction effects, is to stratify samples into discrete subpopulations based on their environmental exposure. Then linear regression (Eqs. 1.1-1.2, see Section 1.1.5) is applied to each strata and subsequently the marginal variant effects are compared to assess whether there is a significant difference in these effects across the different subgroups (in the remainder of this introduction section, I will refer to this method as the ‘stratification interaction test’)<sup>212,216,217</sup>. However, as more detailed environmental data is collected, allowing for finer stratification of the population, these methods are no longer optimal as

the subpopulations become too small to obtain stable estimates of the variant effects<sup>216</sup>.

A second commonly used method to test for interaction effects relies on a simple extension to the linear regression models used for GWAS (described by Eqs. 1.1-1.2, see Section 1.1.5), via the addition of two additional fixed effect terms ( $G \times E$  and  $E$ ), which can be cast as:

$$\text{logit}(\mathbf{y}_D) = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_G + \underbrace{\mathbf{x} \odot \mathbf{e}\beta_{G \times E}}_{G \times E} + \underbrace{\mathbf{e}\beta_E}_E + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (2.1)$$

where  $\mathbf{y}_D$  is an  $N \times 1$  binary phenotype vector, capturing for example disease status<sup>179,218</sup> or:

$$\mathbf{y} = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_G + \underbrace{\mathbf{x} \odot \mathbf{e}\beta_{G \times E}}_{G \times E} + \underbrace{\mathbf{e}\beta_E}_E + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (2.2)$$

where  $\mathbf{y}$  is an  $N \times 1$  quantitative phenotype vector (in the remainder of this introduction section, I will refer to this method as the ‘linear interaction test’)<sup>179</sup>. In both cases,  $N$  is the number of samples used in the analysis,  $\mathbf{W}$  is an  $N \times a$  covariate design matrix and  $\boldsymbol{\alpha}$  is an  $a \times 1$  vector of corresponding effects,  $\mathbf{x}$  is an  $N \times 1$  genotype vector for the focal variant,  $\odot$  denotes element wise multiplication (Hadamard product),  $\mathbf{e}$  is an  $N \times 1$  environmental exposure vector, and  $\beta_G, \beta_{G \times E}$  and  $\beta_E$  are the marginal genetic, interaction and environment effects, respectively.  $\boldsymbol{\epsilon}$  is an  $N \times 1$  noise vector, modelled as random effect following the multivariate normal distribution:

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}_N), \quad (2.3)$$

where  $\mathbf{I}_N$  is the  $N \times N$  identity matrix<sup>179</sup>.

The interaction test, assesses whether  $\beta_{G \times E} \neq 0$ . Formally the interaction hypothesis test is:

$$H_0 : \beta_{G \times E} = 0 \quad (2.4)$$

vs

$$H_1 : \beta_{G \times E} \neq 0. \quad (2.5)$$

Since researchers are often interested in identifying variants associated with a phenotype, whether that be due to marginal or interaction effects and the presence of interaction effects may mask or reduce the observed marginal effect of a variant<sup>177,179,188,218</sup>, it has been proposed that the same model framework can be used to test for association effects whilst allowing for the presence of interaction



effects. This may enable the identification of variants associated with traits that are not be detected through the use of a conventional LM. This is often referred to as a ‘joint association’ test, which tests whether at least one of two variables has a non-zero effect, i. e.  $[\beta_G, \beta_{G \times E}] \neq \mathbf{0}$ <sup>179,218</sup>. Due to the second degree of freedom that is introduced, these joint tests can be slightly less powered than marginal association tests when there is no dependence on the environment<sup>188</sup>.

Similarly to the marginal association model (see Section 1.1.5), different data encodings can be used to make different assumptions regarding the dependence between the genetic variant and environment including dominance, recessiveness and multiplicative effects<sup>188</sup>. However, similarly to considering higher order interactions (see Section 1.3.2) the benefits of considering these alternative models is likely outweighed by the increase in the degrees of freedom of the test or an increased multiple testing burden<sup>180</sup>.

Using these two aforementioned methods, there has been limited success in robustly identifying interaction effects, namely due to the many shortcomings of interaction tests outlined in Section 1.3.2. As well as concerted efforts to increase sample sizes and improve the quality of environmental data collected in an attempt to improve detection of interaction effects, there have been methodological developments in two main directions.

The first are methods that test a subset of the SNPs based on their marginal genetic effects or variance heterogeneity (and in some cases select a subset of the environments based on marginal effects from environment-wide association studies), thereby reducing the multiple testing burden<sup>180</sup>. These marginal effects can be determined based on a previous association study, literature mining or two step procedures that test all variants for marginal effects in the first step, selecting variants to test for interaction effects at the second step; the latter is valid as long as the marginal test and interaction test are independent of one another<sup>175,177,180,193,219–231</sup>.

A key disadvantage of this approach is that interaction effects may be missed when variants or environments have no or weak marginal effects and evidence does exist to demonstrate that interaction effects can occur for variants with no marginal effects<sup>186,188,226,232,233</sup>. In addition, there will be a bias towards increasing knowledge at loci for which information is already available<sup>188,226,232,233</sup>.

Combining interaction tests based on linear models with variant set test based approaches that were originally developed for marginal association tests (see Section

1.2.1), is a second direction of method development. Analogous to the advantages for marginal association tests, these set test based approaches can increase power to detect interactions by aggregating modest and weak interaction effects across variants. In addition, a regional set of genetic markers can better tag an unmeasured causal variant than individual genetic markers, thereby boosting the power to detect interaction effects (see Section 1.3.2 for a description of how variance explained depends on the tagging strength)<sup>234</sup>. Additionally, similarly to the benefit of filtering approaches, these set tests reduce the multiple testing burden. There are a number of set test based approaches for interaction testing, which as for marginal association set tests, can be broadly categorised as burden<sup>207,212,213,235–239</sup>, adaptive burden<sup>240–242</sup>, variance component<sup>243</sup> and combined burden and variance component approaches (see Section 1.2.1)<sup>244</sup>. Further more specific details of interaction set test approaches are provided in Section 2.2.7.

Method development to test for  $G \times E$  effects jointly at multiple environmental variables has largely been ignored. The exception is a study published in 2016 by Young *et al.*<sup>195</sup>, where the authors test for interaction effects with multiple lifestyle-based environment factors on BMI, focussing on a single variant (*rs1421085*) within the *FTO* gene. This analysis is based on a two step procedure, where in the first step they identify environments that display evidence of marginal effects on BMI whilst accounting for the effect of the other considered environments and then in the second step test for interactions using an extended version of the linear interaction test described above.

Briefly the linear interaction test can be modified to test for multi-environment  $G \times E$  through the inclusion of additional interaction and environment terms (see Section 2.3.2 for full details). However, as I will show in Section 2.4.4, these methods are not always calibrated, in particular when the number of environments is large compared to the sample size. In theory, the stratification interaction test can also be modified such that multiple environments are used to stratify the samples into a greater number of subgroups. However, discretisation of continuous environmental variables results in loss of information and in addition, subgrouping of samples reduces the sample size for fitting each LM, both of which result in a loss of potential power to identify  $G \times E$  effects.

Therefore, in this chapter, I will present a multi-environment  $G \times E$  test, the structured linear mixed model (StructLMM), that generalises and overcomes the problems of existing interaction tests. Whilst primarily designed to identify interaction effects, I will show that the framework can also be used to test for joint

associations, in a similar vein to Kraft *et al.*<sup>218</sup> and can be used for downstream interpretation (see Chapter 3). The method has been implemented such that it is computationally efficient, enabling analysis of cohorts containing hundreds of thousands of individuals and hundreds of environments. StructLMM is flexible, such that included environments can be continuous and/or binary, external (e.g. lifestyle factors) or intrinsic (e.g. tissue or cell type) in nature, or be genetic variants themselves (i.e. testing for epistatic interactions). StructLMM is freely available from <https://github.com/limix/struct-lmm> and is supported within the LIMIX framework<sup>78</sup> at <https://github.com/limix/limix>. For tutorials and illustrations on how to use the model, see <http://struct-lmm.readthedocs.io>.

In Section 2.2, I describe the StructLMM model, in Section 2.3, I provide details of other models that are used to benchmark StructLMM and in Section 2.4, I present results from simulation experiments.

Some of the material presented in this chapter is joint work with Francesco Paolo Casale. Specifically, whilst I derived the first version of the mathematics underpinning the method described in Section 2.2.5, suggestions from Francesco Paolo Casale to improve the elegance of the proof, were incorporated. In addition, Francesco Paolo Casale suggested possible ways in which StructLMM may be made more computationally efficient that I subsequently incorporated (see Section 2.2.6). Finally, the design of the simulation experiments and production of simulation results described in Section 2.4 was joint work. This work has been published by Nature Genetics<sup>245</sup>. A copy of this publication can be found in Appendix A (apart from the Supplementary Tables which are available at <https://www.nature.com/articles/s41588-018-0271-0>).

## 2.2 StructLMM

In this section, I introduce the StructLMM model and provide an overview of the available interaction and joint association tests, followed by assumptions that are made throughout this work. I then provide details of the testing procedures, including the generation of the environment covariance matrix and the mathematics underpinning the statistical tests before describing specific implementation details and the corresponding computational complexities. Finally, I place StructLMM in context to existing interaction and joint association tests, describing technical similarities.

### 2.2.1 The model

**LMM** As described in Section 1.1.6 (Eq. 1.6), a conventional LMM used to test for associations between a  $N \times 1$  phenotype vector,  $\mathbf{y}$  (where  $N$  is the sample size of the analysis), and a  $N \times 1$  genotype vector of a focal variant,  $\mathbf{x}$  can be written as:

$$\mathbf{y} = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{\mathbf{u}}_{\text{Confounding}} + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (2.6)$$

where the scalar  $\beta_G$  is the focal variant effect,  $\mathbf{W}$  is an  $N \times a$  design matrix of  $a$  covariates and  $\boldsymbol{\alpha}$  is an  $a \times 1$  vector of the corresponding effects,  $\mathbf{u}$  is an  $N \times 1$  vector that can account for population structure, other confounding factors (between  $\mathbf{y}$  and  $\mathbf{x}$ ) or additive environment effects, and  $\boldsymbol{\epsilon}$  is an  $N \times 1$  noise vector. The final two terms are modelled as random effects following multivariate normal distributions:

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \mathbf{R}), \quad (2.7)$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}_N), \quad (2.8)$$

where  $\mathbf{R}$  is an  $N \times N$  matrix and  $\mathbf{I}_N$  is the  $N \times N$  identity matrix.

An association is present when  $\beta_G \neq 0$  (formal definitions of this hypothesis test, are provided in Section 1.1.5). This association test is performed under the assumption that the effect of a variant is the same across all  $N$  samples (referred to as a persistent genetic effect), whilst in reality the effect of the variant may vary from one individual to the next due to differences in their environmental exposures (G×E).

**StructLMM** The StructLMM model builds on the conventional LMM (Eq. 2.6), allowing for heterogeneity in variant effect sizes across the  $N$  individuals due to G×E through the inclusion of an additional term,  $\mathbf{x} \odot \boldsymbol{\beta}_{G \times E}$ , resulting in the following model:

$$\mathbf{y} = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{\mathbf{x} \odot \boldsymbol{\beta}_{G \times E}}_{\text{G} \times \text{E}} + \underbrace{\mathbf{u}}_{\text{E}} + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (2.9)$$

where  $\odot$  denotes element wise multiplication (Hadamard product), such that  $\mathbf{x} \odot \boldsymbol{\beta}_{G \times E}$  is an  $N \times 1$  vector, and is equivalent to  $\text{diag}(\mathbf{x})\boldsymbol{\beta}_{G \times E}$  where  $\text{diag}(\mathbf{x})$  is an  $N \times N$  matrix which has non-zero values only on the diagonal.  $\boldsymbol{\beta}_{G \times E}$  is an  $N \times 1$  vector, modelled as a random effect, following the multivariate normal distribution:

$$\boldsymbol{\beta}_{G \times E} \sim \mathcal{N}(\mathbf{0}, \sigma_{G \times E}^2 \boldsymbol{\Sigma}), \quad (2.10)$$

where  $\Sigma$  is  $N \times N$  symmetric matrix that encodes the environmental similarity between pairs of individuals (see Section 2.2.4 for details). In the special case, where  $\beta_{G \times E} = \mathbf{0}$  or alternatively,  $\sigma_{G \times E}^2 = 0$ , Eq. 2.9 is identical to the conventional LMM described in Eq. 2.6. This form of the model is used to obtain the marginal model (see Section 2.2.5) of the StructLMM interaction test and downstream interpretation tools described in Chapter 3 (see Section 2.2.5). I will refer to this marginal form of the model as the ‘interaction test marginal distribution’.

If I also model, the persistent genetic effect,  $\beta_G$ , as a random effect, following the normal distribution:

$$\beta_G \sim \mathcal{N}(0, \sigma_G^2), \quad (2.11)$$

it is possible to rewrite Eq. 2.9<sup>†</sup>, such that the persistent,  $\beta_G$ , and heterogeneous,  $\beta_{G \times E}$ , genetic effect terms are captured by a single term,  $\beta$  as follows:

$$\mathbf{y} = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x} \odot \boldsymbol{\beta}}_{G+G \times E} + \underbrace{\mathbf{u}}_E + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (2.12)$$

where  $\beta$  is an  $N \times 1$  vector, modelled as a random effect, following the multivariate distribution:

$$\beta \sim \mathcal{N}(\mathbf{0}, \sigma_G^2 \mathbf{1}_N \mathbf{1}_N^T + \sigma_{G \times E}^2 \Sigma), \quad (2.13)$$

where  $\mathbf{1}_N$  is an  $N \times 1$  vector of ones such that  $\mathbf{1}_N \mathbf{1}_N^T$  is an  $N \times N$  matrix of ones. The first covariance term models full correlation of the SNP effect across individuals, thereby capturing the component of the SNP effect that is persistent across all individuals and the second covariance term is identical to that described in Eq. 2.10, and thus allows the SNP effect to vary across individuals due to differences in environmental exposures. This form of the model is used to obtain the marginal model of the StructLMM joint association test (see Section 2.2.5). I will refer to this marginal form of the model as the ‘association test marginal distribution’.

With these random effect designs that capture  $G \times E$  through a single covariance term,  $\sigma_{G \times E}^2 \Sigma$  (see Eqs. 2.10 and 2.13), both forms of the StructLMM model (described in Eqs. 2.9 and 2.12) can be used to model different possible  $G \times E$  scenarios, thereby generalising previous approaches. In the simplest case, to assess  $G \times E$  driven by two subgroups of individuals (Fig. 2.1a), for example gender specific effects or by stratifying the individuals based on their environmental exposure (e. g. highly and lowly exposed), a block diagonal covariance matrix,  $\Sigma$ , can capture this (Fig. 2.1d). If the number of subgroups considered is increased, either by

---

<sup>†</sup>noting that  $\mathbf{x}\beta_G = \mathbf{x} \odot \boldsymbol{\beta}_G$ , where  $\beta_G$  and  $\boldsymbol{\beta}_G$  are modelled as a random effects following the distributions  $\beta_G \sim \mathcal{N}(0, \sigma_G^2)$  and  $\boldsymbol{\beta}_G \sim \mathcal{N}(\mathbf{0}, \sigma_g^2 \mathbf{1}_N \mathbf{1}_N^T)$ , respectively.

increasing the number of environmental variables that are used for stratification, for example, four subgroups could be defined by combining the aforementioned gender and environmental exposure levels or by stratifying the data into a greater number of subgroups when non-binary environmental variables are available (Fig. 2.1b), a block diagonal covariance matrix,  $\Sigma$ , with greater rank can be used. In many cases, it will be possible to define a hierarchy over these different environmental subgroups, which can be captured by off diagonal terms in the covariance matrix,  $\Sigma$  (Fig. 2.1e). In the limiting case, where many environments or a continuous exposure variable are assessed, each individual is its own subgroup, resulting in per-individuals effect sizes (Fig. 2.1c, f).

## 2.2.2 StructLMM interaction and association tests

The StructLMM framework allows for performing two different tests:

1. The first is an interaction test, that tests whether the effect of a variant is significantly different under different environmental exposures, whilst accounting for persistent genetic effects under the null. This is equivalent to testing whether  $\sigma_{G \times E}^2 \neq 0$  for both forms of the model given by Eqs. 2.9-2.13.

As outlined in Section 1.1.5, there are slightly different statistical testing procedures that can be used to perform this interaction test. Since the score test is both computationally efficient and the framework readily adapted for the joint association test (where the score test yields even greater efficiency advantages), I employ a score-based test.

The formal interaction hypothesis test is:

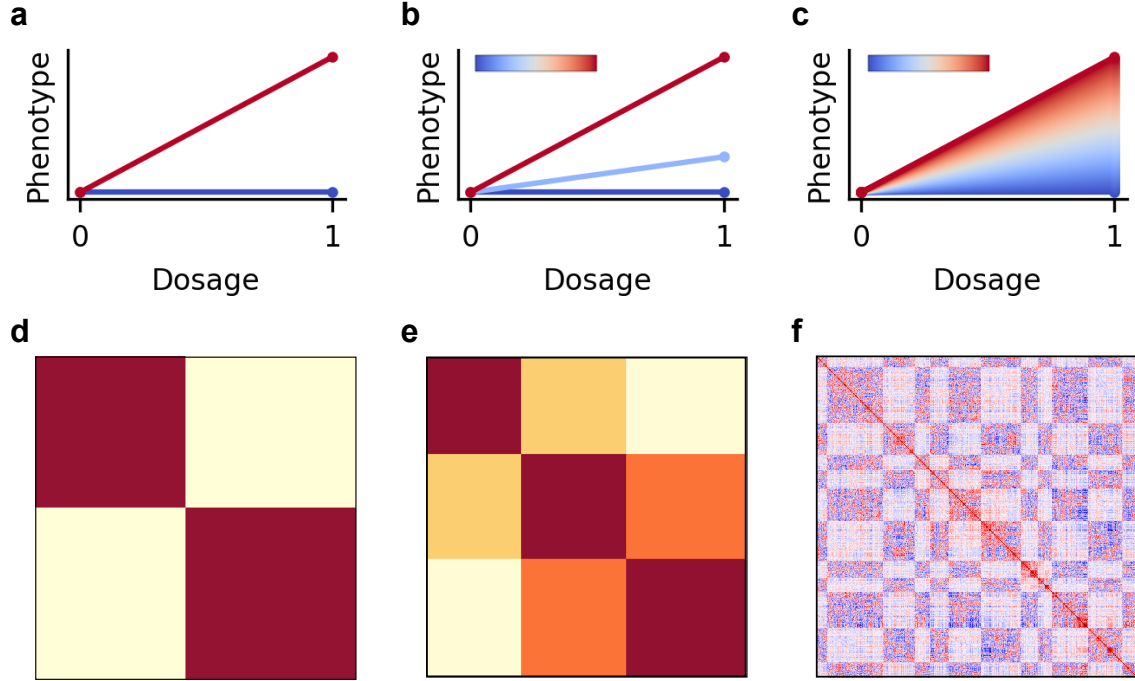
$$H_0 : \sigma_{G \times E}^2 = 0 \quad (2.14)$$

vs

$$H_1 : \sigma_{G \times E}^2 > 0, \quad (2.15)$$

noting that  $\sigma_{G \times E}^2$  is a variance parameter and hence can only take positive values.

As a result, the tested parameter is on the boundary of the parameter space under the null model<sup>246–249</sup>, such that the score test statistic does not follow the usual  $\chi_1^2$  but instead a mixture of  $\chi^2$  distributions. Since, in general,  $\Sigma$  is not an identity matrix or block diagonal, the score test statistic will not



**Fig. 2.1 Different  $G \times E$  scenarios that can be tested using StructLMM and corresponding examples of environment covariance structures,  $\Sigma$  |**  
 (a) Basic  $G \times E$  with two subgroups (one subgroup represented by the blue line and the other by the red line) of individuals stratified based on their environmental exposures. Deviation in the difference of the average phenotypic value between the two subgroup across the two genotype dosages signifies  $G \times E$  effects are present. (b)  $G \times E$  with multiple (three in this illustrative example) subgroups present due to the availability of multiple environmental exposure variables or due to stratification of the environmental data into a greater number of subgroups. Subgroup hierarchy based on environmental exposure levels may be present. (c)  $G \times E$  in the limiting case where effect sizes are defined per-individual, which can be driven both by a large number of environmental variables and through the use of continuous measures of environmental exposure. (d-f) Examples of environment covariance structures,  $\Sigma$  for different settings, (d) when there exist two subgroups, (e) when there exist multiple (three in this illustrative example) subgroups, in the presence of subgroup hierarchy and (f) when there exist many binary and continuous environmental exposures.

follow the conventional 50:50 mixture of  $\chi_0^2$  and  $\chi_1^{250-252}$  but instead has been shown to follow a mixture of  $\chi_1^2$  variables, weighted based on the structure of  $\Sigma^{118,123,251,253}$  (see Section 2.2.5 for details). Full details of the interaction testing procedure are described in Section 2.2.5.

2. Whilst primarily designed for interaction testing, the framework can also be used for joint association testing, similar to Kraft *et al.*<sup>218</sup>, testing for associations between a variant and a phenotype, accounting for possible heterogeneity in the variant effect due to G×E. More explicitly, this corresponds to testing whether  $[\beta_G, \sigma_{G \times E}^2] \neq \mathbf{0}$  in Eqs. 2.9-2.10, or alternatively,  $[\sigma_G^2, \sigma_{G \times E}^2] \neq \mathbf{0}$  in the form of the model described by Eqs. 2.12-2.13. In the latter case, the formal joint association hypothesis test is:

$$H_0 : [\sigma_G^2, \sigma_{G \times E}^2] = \mathbf{0} \quad (2.16)$$

vs

$$H_1 : [\sigma_G^2, \sigma_{G \times E}^2] > \mathbf{0}, \quad (2.17)$$

again noting that  $\sigma_G^2$  and  $\sigma_{G \times E}^2$  are variance parameters and hence can only take positive values.

Again, for computational efficiency a score-based test is used; for the joint association test, it can be seen that the score-based implementation yields even greater efficiency gains compared to the LRT and Wald test than for the interaction test, since a single null model needs to be fit to perform a genome-wide scan.

Again, the tested parameters lie on the boundary of the parameter space and deriving the appropriate mixture of  $\chi^2$  variables is substantially more difficult in this setting<sup>249</sup>, where two variance components are tested, than in the interaction test setting. To avoid computationally inefficient Monte Carlo based<sup>254,255</sup> and permutation based solutions<sup>249</sup>, StructLMM follows the approach taken in SKAT-O<sup>123</sup>, where the variance component test of two free parameters is reduced to a variance component test with a single free parameter. This is achieved via the introduction of the parameter  $\rho$ , which I grid search over (i. e. a test for each value of  $\rho$  is conducted, which is followed by an adjustment to account for this ‘multiple testing’, resulting in a single final P value per variant; see Section 2.2.5 for details) such that Eq. 2.13 can be rewritten as:

$$\beta \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{tot}}^2[(1 - \rho)\mathbf{1}_N\mathbf{1}_N^T + \rho\Sigma]). \quad (2.18)$$



By comparison with Eq. 2.13, it can be seen that  $\sigma_{\text{tot}}^2(1 - \rho) \equiv \sigma_G^2$  and  $\sigma_{\text{tot}}^2\rho \equiv \sigma_{G \times E}^2$ , from which it follows that  $\sigma_{\text{tot}}^2 \equiv \sigma_G^2 + \sigma_{G \times E}^2$ <sup>†</sup>.  $\sigma_{\text{tot}}^2$  denotes the total variance of the genetic effect (encompassing both the persistent and  $G \times E$  components), after appropriate adjustment for the matrix variance of  $\Sigma$  (see Section 3.2.1 for full details), whilst the parameter  $\rho$  corresponds to the fraction of the total genetic variance that can be explained by  $G \times E$ :

$$\rho = \frac{\sigma_{G \times E}^2}{(\sigma_G^2 + \sigma_{G \times E}^2)} = \frac{\sigma_{G \times E}^2}{\sigma_{\text{tot}}^2} \quad (2.19)$$

The formal hypothesis test is now:

$$H_0 : \sigma_{\text{tot}}^2 = 0 \quad (2.20)$$

vs

$$H_1 : \sigma_{\text{tot}}^2 > 0 \quad (2.21)$$

Full details of the joint association test are described in Section 2.2.5.

### 2.2.3 Simplifying assumptions

I now make three simplifying assumptions that hold throughout the rest of this thesis.

1. **The individuals included in the analysis are unrelated.** In theory, StructLMM can account for confounding due to population structure using a random effect term (as described in Eqs. 2.6 and 2.7), where  $\mathbf{R}$  is defined as the genetic relatedness matrix (see Section 1.1.6), such that related individuals can be included in the analysis. However, the genetic relatedness matrix,  $\mathbf{R}$ , is typically full rank and hence StructLMM would not scale linearly with the number of samples. To retain computational efficiency, such that the method can be applied to large cohorts, I account for population structure using principal components, which will be included as covariates (see Section 1.1.6).
2. **The covariance matrix used to account for additive environment effects is identical to the covariance matrix used to model  $G \times E$ .** Whilst in general, different covariance structures can be used to model the additive environment effects (which accounts for direct correlation between the

---

<sup>†</sup>Substituting  $\sigma_{\text{tot}}^2\rho \equiv \sigma_{G \times E}^2$  into  $\sigma_{\text{tot}}^2(1 - \rho) \equiv \sigma_G^2$ , gives  $\sigma_{\text{tot}}^2 - \sigma_{G \times E}^2 \equiv \sigma_G^2$ , which once rearranged gives,  $\sigma_{\text{tot}}^2 \equiv \sigma_G^2 + \sigma_{G \times E}^2$

environments and the phenotype) and the interaction effects, for simplicity, I assume they are the identical. Explicitly, the additive environment effect is modelled through the random effect term,  $\mathbf{u}$ , with  $\mathbf{R} = \mathbf{\Sigma}$  (Eqs. 2.7 and 2.9).

3. **The linear covariance function is used to build the environment covariance,  $\mathbf{\Sigma}$ .** Whilst in principle, any valid covariance function<sup>256</sup> can be used to define the environmental similarity matrix  $\mathbf{\Sigma}$ , in this thesis I consider only the linear covariance function, defined as:

$$\mathbf{\Sigma} = \mathbf{E}\mathbf{E}^T, \quad (2.22)$$

where  $\mathbf{E}$  is an  $N \times L$  matrix, where  $L$  is the number of tested environments (which may be continuous and/or discrete) and

$$\mathbf{E} = f(\mathbf{E}_{\text{raw}}), \quad (2.23)$$

where  $f$  represent any functional mapping (equivalent to using basis functions in linear regression) from the raw observed environment covariates,  $\mathbf{E}_{\text{raw}}$ , to the environment covariates,  $\mathbf{E}$ , used to build the covariance matrix,  $\mathbf{\Sigma}$  (see Section 2.2.4 for descriptions of commonly used functional mappings).

The use of a linear covariance function was motivated by two appealing properties. The first is that in general the number of environments,  $L$ , will be much smaller than the number of individuals,  $N$  ( $L \ll N$ ), meaning that the linear covariance function will be low rank. This enables algebraic reformulation, which in turn results in computational complexity that scales linearly with the number of samples,  $N$ , i.e.  $O(N)$  complexity (see Section 2.2.6 for details). The second is that a linear covariance is directly interpretable as there is a one-to-one correspondence between StructLMM and multivariate linear regression using  $L$  covariates to capture the interaction term (see Section 2.2.4).

## 2.2.4 Environment covariance, $\mathbf{\Sigma}$

As already outlined in the previous section, in this work, I use the linear covariance function to build an environmental similarity matrix,  $\mathbf{\Sigma}$ . In this section, I will describe commonly employed functional mappings (see Eq. 2.23) and the broad range of settings, including non-linear relationships that a linear covariance function can capture. Finally, I will provide details of the one-to-one correspondence between

multivariate linear regression and StructLMM, when a linear covariance function is used to capture environmental similarity.

### Common functional mappings

A commonly employed functional mapping, is to normalise the  $L$  environmental variables such that each exposure has mean 0 and standard deviation 1. A standard strategy is to then rescale these normalised features by a factor of  $\frac{1}{\sqrt{L}}$  such that the resulting covariance matrix,  $\Sigma$  has sample mean 0 and sample variance 1<sup>257</sup>. Other valid strategies, include row standardisation such that either the resulting covariance matrix,  $\Sigma$  has per-individual variance 1 or is a correlation matrix<sup>256</sup>. These different standardisation procedures make slightly different assumptions regarding the variance explained by different environments within the interaction and additive environment terms, similar to building linear covariances using genetic factors<sup>74</sup>.

### Settings captured by linear covariance function

The linear covariance function is able to capture different interaction settings. This includes group specific effects through the use of  $L$  binary environmental variables, where  $L$  is the number of defined subgroups, resulting in a block diagonal matrix (Fig. 2.1d) or if some of these  $L$  environmental variables are ordinal then a natural hierarchy amongst the different subgroups can be captured resulting in off-diagonal non-zero values (Fig. 2.1e) and finally if some of the  $L$  exposures are continuous then per-individual variant effects can be modelled (Fig. 2.1f). Despite the name, a linear covariance function can also capture non-linear relationships between the observed environments, through the introduction of additional transformed environmental variables, similar to using basis functions for non-linear regression<sup>256</sup>. This may include exponentiation of the observed environments or combining several variables. For example, the  $L$  environmental variables,  $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_L\}$  can be combined with a categorical variable,  $c \in \{0, 1\}^{N \times 1}$ , such as gender, to define  $\mathbf{E} = [\mathbf{c} \odot \mathbf{e}_1, \mathbf{c} \odot \mathbf{e}_2, \dots, \mathbf{c} \odot \mathbf{e}_L, (1 - \mathbf{c}) \odot \mathbf{e}_1, (1 - \mathbf{c}) \odot \mathbf{e}_2, \dots, (1 - \mathbf{c}) \odot \mathbf{e}_L]$ , an  $N \times 2L$  matrix. This is the approach taken to define gender and age adjusted environmental variables (Section 2.4.1 and Chapters 3 and 4).

## Derivation as marginalised linear $\mathbf{G} \times \mathbf{E}$ interaction model

Using the linear covariance function to define the environmental similarity matrix,  $\Sigma$ , facilitates direct interpretation, since StructLMM can be derived from a multivariate linear interaction model, analogous to the well-known relationship between multivariate Bayesian linear regression and LMMs (as outlined in Section 2.2.3, simplifying assumption 3)<sup>75</sup>. Denoting the  $N \times 1$ ,  $L$  environment vectors as  $\mathbf{e}_1, \dots, \mathbf{e}_L$ , a multivariate linear interaction model can be cast as:

$$\mathbf{y} = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{covariates}} + \underbrace{\mathbf{x}\beta_{\mathbf{G}}}_{\mathbf{G}} + \underbrace{\sum_{l=1}^L (\mathbf{x} \odot \mathbf{e}_l)\beta_{\mathbf{G} \times \mathbf{E}_l}}_{\mathbf{G} \times \mathbf{E}} + \underbrace{\sum_{l=1}^L \mathbf{e}_l\beta_{\mathbf{E}_l}}_{\mathbf{E}} + \underbrace{\boldsymbol{\epsilon}}_{\text{noise}}. \quad (2.24)$$

By defining the following priors on the variance of  $\beta_{\mathbf{G} \times \mathbf{E}_l}$  and  $\beta_{\mathbf{E}_l}$ :

$$\beta_{\mathbf{G} \times \mathbf{E}_l} \sim \mathcal{N}(0, \sigma_{\mathbf{G} \times \mathbf{E}}^2), \quad (2.25)$$

$$\beta_{\mathbf{E}_l} \sim \mathcal{N}(0, \sigma_e^2) \quad (2.26)$$

and marginalising over  $\beta_{\mathbf{G} \times \mathbf{E}_l}$  and  $\beta_{\mathbf{E}_l}$ , it can be seen that this multi-environment linear interaction model is equivalent to the form of StructLMM described by Eqs. 2.7-2.10. The multivariate linear interaction model described here, is one of the comparison partners (Section 2.3) used in simulation experiments (Section 2.4) and in application to real data (Chapter 3).

### 2.2.5 Statistical testing

In this section, I will describe in detail the statistical testing procedures and the underpinning mathematics for both the interaction and joint association StructLMM tests (outlined in Section 2.2.2). I will first describe the marginal models used for parameter inference, followed by the different steps required for the statistical tests.

## Marginal models

A marginalised form of Eq. 2.9 can be obtained by integrating over  $\beta_{G \times E}^\dagger$  (Eq. 2.10) and  $\mathbf{u}$  (Eq. 2.7), such that:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{W}\boldsymbol{\alpha} + \mathbf{x}\beta_G, \sigma_{G \times E}^2 \text{diag}(\mathbf{x})\boldsymbol{\Sigma}\text{diag}(\mathbf{x}) + \sigma_e^2 \boldsymbol{\Sigma} + \sigma_n^2 \mathbf{I}_N). \quad (2.27)$$

This marginalised form of the model is used for the interaction test and for the downstream analysis described in Chapter 3 and is referred to as the ‘interaction test marginal distribution’.

A fully marginalised form of the model can be obtained by additionally integrating over the persistent genetic effect. This is equivalent to integrating over  $\beta$  (Eq. 2.18) and  $\mathbf{u}$  (Eq. 2.7) using the form of the model described in Eq. 2.12, such that:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{W}\boldsymbol{\alpha}, \sigma_{\text{tot}}^2 \mathbf{K}_\rho + \sigma_e^2 \boldsymbol{\Sigma} + \sigma_n^2 \mathbf{I}_N), \quad (2.28)$$

where:

$$\mathbf{K}_\rho = (1 - \rho) \text{diag}(\mathbf{x}) \mathbf{1}_N \mathbf{1}_N^T \text{diag}(\mathbf{x}) + \rho \text{diag}(\mathbf{x}) \boldsymbol{\Sigma} \text{diag}(\mathbf{x}). \quad (2.29)$$

Noting that  $\text{diag}(\mathbf{x}) \mathbf{1}_N \equiv \mathbf{x}$ ,  $\mathbf{K}_\rho$  can be simplified to:

$$\mathbf{K}_\rho = (1 - \rho) \mathbf{x} \mathbf{x}^T + \rho \text{diag}(\mathbf{x}) \boldsymbol{\Sigma} \text{diag}(\mathbf{x}). \quad (2.30)$$

This marginalised form of the model is used for the joint association test and is referred to as the ‘association test marginal distribution’.

## Fitting the models under the null

The score test requires fitting the models under the null hypothesis. For the StructLMM interaction test, the null model is:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{x}\beta_G + \mathbf{u} + \boldsymbol{\epsilon}, \quad (2.31)$$

where the interaction test marginal distribution (based on Eq. 2.27):

$$\mathbf{y} \sim \mathcal{N}(\mathbf{W}\boldsymbol{\alpha} + \mathbf{x}\beta_G, \sigma_e^2 \boldsymbol{\Sigma} + \sigma_n^2 \mathbf{I}_N), \quad (2.32)$$

is used for parameter inference.

---

<sup>†</sup>using  $\mathbf{x} \odot \beta_{G \times E} \equiv \text{diag}(\mathbf{x}) \beta_{G \times E}$

For the StructLMM joint association test, the null model is:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{u} + \boldsymbol{\epsilon}, \quad (2.33)$$

where the association test marginal distribution (based on Eq. 2.28):

$$\mathbf{y} \sim \mathcal{N}(\mathbf{W}\boldsymbol{\alpha}, \sigma_e^2 \boldsymbol{\Sigma} + \sigma_n^2 \mathbf{I}_N), \quad (2.34)$$

is used for parameter inference.

Parameter inference is performed using a previously implemented REML-based framework, LIMIX<sup>78</sup>, to identify the optimal parameters  $\hat{\boldsymbol{\alpha}}$ ,  $\hat{\sigma}_e^2$  and  $\hat{\sigma}_n^2$  and in addition  $\hat{\beta}_G$  for the interaction test.

Subsequently the covariance under the null,  $\mathbf{K}_0$ , for both the interaction and joint association test is:

$$\mathbf{K}_0 = \hat{\sigma}_e^2 \boldsymbol{\Sigma} + \hat{\sigma}_n^2 \mathbf{I}_N \quad (2.35)$$

### Score-based test statistic

As described in Lippert *et al.*<sup>253</sup>, the score-based test statistic,  $Q$ , is given by:

$$Q = \frac{1}{2} \mathbf{y}^T \mathbf{P}_0 \frac{\partial \mathbf{K}}{\partial \theta} \mathbf{P}_0 \mathbf{y} \quad (2.36)$$

where  $\mathbf{K}$  is the covariance of the marginal likelihood (see Eq. 2.27 and Eq. 2.28 for the interaction and joint association tests, respectively),  $\theta$  is the focal parameter of the hypothesis test ( $\sigma_{G \times E}^2$ , Eqs. 2.14-2.15 and  $\sigma_{\text{tot}}^2$ , Eqs. 2.20-2.21 for the interaction and joint association tests, respectively) and:

$$\mathbf{P}_0 = \mathbf{K}_0^{-1} - \mathbf{K}_0^{-1} \mathbf{X} (\mathbf{X}^T \mathbf{K}_0^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{K}_0^{-1}, \quad (2.37)$$

where  $\mathbf{K}_0$  is the covariance of the marginal likelihood under the null hypothesis (defined in Eq. 2.35) and  $\mathbf{X}$  is the combined design matrix of all fixed effects in the marginal likelihood, such that  $\mathbf{X} = [\mathbf{W}, \mathbf{x}]$  and  $\mathbf{X} = \mathbf{W}$  for the interaction and joint association test, respectively.

**Derivative of the covariance,  $\mathbf{K}$ , w.r.t. the parameter of interest,  $\theta$ :**  
 $\frac{\partial \mathbf{K}}{\partial \theta}$

For the interaction test, using the marginal likelihood described by Eq. 2.27, where  $\mathbf{K} = \sigma_{\text{G} \times \text{E}}^2 \text{diag}(\mathbf{x}) \boldsymbol{\Sigma} \text{diag}(\mathbf{x}) + \sigma_e^2 \boldsymbol{\Sigma} + \sigma_n^2 \mathbf{I}_N$  and  $\theta = \sigma_{\text{G} \times \text{E}}^2$ ,

$$\frac{\partial \mathbf{K}}{\partial \sigma_{\text{G} \times \text{E}}^2} = \text{diag}(\mathbf{x}) \boldsymbol{\Sigma} \text{diag}(\mathbf{x}). \quad (2.38)$$

For the joint association test, using the marginal likelihood described by Eq. 2.28, where  $\mathbf{K} = \sigma_{\text{tot}}^2 \mathbf{K}_\rho + \sigma_e^2 \boldsymbol{\Sigma} + \sigma_n^2 \mathbf{I}_N$  and  $\theta = \sigma_{\text{tot}}^2$ ,

$$\frac{\partial \mathbf{K}}{\partial \sigma_{\text{tot}}^2} = \mathbf{K}_\rho, \quad (2.39)$$

where  $\mathbf{K}_\rho = (1 - \rho) \mathbf{x} \mathbf{x}^T + \rho \text{diag}(\mathbf{x}) \boldsymbol{\Sigma} \text{diag}(\mathbf{x})$  as defined in Eq. 2.30.

It can be seen that the derivative for the interaction test, is the special case of the derivative of the association test when  $\rho = 1$ , i.e.  $\mathbf{K}_1 = \text{diag}(\mathbf{x}) \boldsymbol{\Sigma} \text{diag}(\mathbf{x})$  (I use the notation  $\mathbf{K}_1$  throughout the rest of this chapter to denote this interaction test derivative).

### Score-based test statistic for interaction test

The score-based test statistic for the interaction test is thus:

$$Q_{\text{int}} = \frac{1}{2} \mathbf{y}^T \mathbf{P}_0 \mathbf{K}_1 \mathbf{P}_0 \mathbf{y}, \quad (2.40)$$

where  $\mathbf{K}_1$  is as defined above.

### Evaluating the significance of the score-based interaction test statistic

It can be shown that under the null, the score-based test statistic,  $Q$ , follows a weighted sum of  $\chi_1^2$  variables<sup>253</sup>:

$$Q \sim \sum_k \lambda_k \chi_1^2, \quad (2.41)$$

where  $\lambda_k$  are the non-zero eigenvalues of  $\frac{1}{2} \mathbf{P}_0^{\frac{T}{2}} \frac{\partial \mathbf{K}}{\partial \theta} \mathbf{P}_0^{\frac{1}{2}}$ .

There are a number of methods that can be used to evaluate the significance of

the score-based test statistic,  $Q_{\text{int}}$ , given the null distribution (I refer the reader to Duchesne *et al.*<sup>258</sup> and Wu *et al.*<sup>259</sup> for further details). For the interaction test, I follow the approach taken in SKAT<sup>118</sup>, where P values are calculated using Davies method<sup>258,260</sup>, an exact characteristic inversion method, switching to the modified moment matching approximation method<sup>123,258,261†</sup> when this fails to converge.

## Test statistic for the joint association test

The score-based test statistic for the joint association test is:

$$Q_{\rho} = \frac{1}{2} \mathbf{y}^T \mathbf{P}_0 \mathbf{K}_{\rho} \mathbf{P}_0 \mathbf{y}, \quad (2.42)$$

where  $\mathbf{K}_{\rho}$  is as defined in Eq. 2.30.

However, this requires setting  $\rho$  to a given value but in general this parameter is unknown. I therefore follow a similar approach to that taken in SKAT-O<sup>123</sup>, which performs a grid search over  $\rho$ . Explicitly, this involves:

1. Define a set,  $r = 1, \dots, R$ , of values for  $\rho$  to grid search over,  $0 = \rho_1 < \rho_2 < \dots < \rho_R = 1$  (In all analyses, I use [0.0, 0.5, 0.75, 0.84, 0.91, 0.96, 0.99, 1.0] by default).
2. Calculate the score-based test statistic,  $Q_{\rho_r}$  for each  $\rho_r$  (using Eq. 2.42). For each of these test statistics a P value,  $P_{\rho_r}$ , is calculated; this is achieved following the same process as was used to obtain the interaction test P value, namely that these score-based test statistics follow a weighted sum of  $\chi_1^2$  variables under the null (Eq. 2.41). In this step the modified moment matching approximation method<sup>123,258,261</sup> is used since it is computationally more efficient than the exact Davies method<sup>258,259</sup> and as will become clear in the following steps these P values are not the final ones reported.
3. Calculate the test statistic  $T = \min\{P_{\rho_1}, P_{\rho_2}, \dots, P_{\rho_R}\}$ .
4. Compute the P value of the test statistic,  $T$ . Details on how this is done are provided below.

---

<sup>†</sup>The moment-matching method is an approximation method that is based on matching the first three moments (mean, variance and skewness) but can lead to inflated Type 1 error rates, in particular for small P values<sup>259</sup>. A modified version of this moment-matching method was implemented in SKAT, which matches the kurtosis instead of skewness, to improve the tail probability approximation<sup>123,258,261</sup> but can still lead to inflated Type 1 error rates, in particular for small P values.



## Evaluating the significance of the test statistic $T$

I now describe how the P value of the test statistic,  $T$  is calculated. The obtained P value describes the significance of the StructLMM association test. The following derivation generalises that required for SKAT-O<sup>123</sup>, where  $\mathbf{K}_\rho$  is the sum of a  $\mathbf{11}^T$  (matrix of ones) and  $\mathbf{I}$  (an identity matrix) matrix. The derivation provided here facilitates the use of a non-identity matrix as the second term in  $\mathbf{K}_\rho$  (Eq. 2.30, see Section 2.2.7 for a comparison to SKAT-O). To the best of my knowledge, this is the first time that explicit derivations have been presented (i.e. explicit derivations are not presented in previous SKAT-O publications).

By definition the P value, is the probability under the null hypothesis of obtaining a test statistic,  $t$ , at least as extreme as the one observed,  $T$ :

$$\begin{aligned} P &= \Pr(t < T) \\ &= \Pr(\min\{p_{\rho_1}, p_{\rho_2}, \dots, p_{\rho_R}\} < T) \\ &= 1 - \Pr(p_{\rho_1} \geq T, p_{\rho_2} \geq T, \dots, p_{\rho_R} \geq T), \end{aligned} \quad (2.43)$$

where  $p_{\rho_r}$  is defined as the  $r^{th}$  P value obtained under the null hypothesis. Again, by definition, the P value,  $p_{\rho_r}$  is the probability under the null hypothesis of obtaining a value,  $q_{\rho_r}$ , at least as extreme as the score-based test statistics  $Q_{\rho_r}$ :

$$p_{\rho_r} = \Pr(Q_{\rho_r} < q_{\rho_r}). \quad (2.44)$$

Substituting, Eq. 2.44 into Eq. 2.43, gives:

$$P = 1 - \Pr(\Pr(Q_{\rho_1} < q_{\rho_1}) \geq T, \Pr(Q_{\rho_2} < q_{\rho_2}) \geq T, \dots, \Pr(Q_{\rho_R} < q_{\rho_R}) \geq T). \quad (2.45)$$

Eq. 2.45 can be rewritten as<sup>†</sup>:

$$P = 1 - \Pr(1 - \Pr(Q_{\rho_1} \geq q_{\rho_1}) \geq T, 1 - \Pr(Q_{\rho_2} \geq q_{\rho_2}) \geq T, \dots, 1 - \Pr(Q_{\rho_R} \geq q_{\rho_R}) \geq T), \quad (2.46)$$

which rearranging gives:

$$P = 1 - \Pr(\Pr(Q_{\rho_1} \geq q_{\rho_1}) \leq 1 - T, \Pr(Q_{\rho_2} \geq q_{\rho_2}) \leq 1 - T, \dots, \Pr(Q_{\rho_R} \geq q_{\rho_R}) \leq 1 - T). \quad (2.47)$$

Either by quantile function definitions or through visualisation of the probability

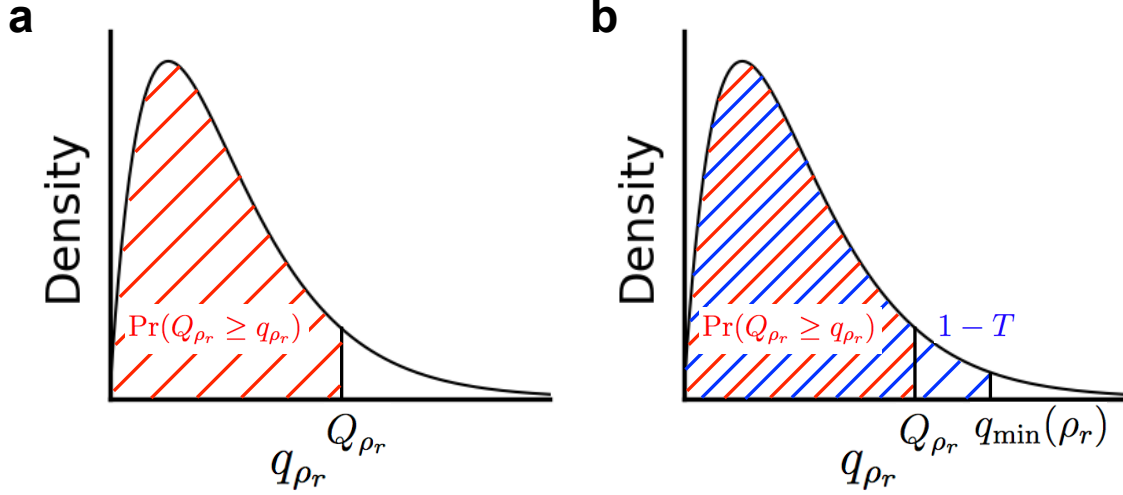
---

<sup>†</sup>using  $\Pr(Q_{\rho_r} < q_{\rho_r}) = 1 - \Pr(Q_{\rho_r} \geq q_{\rho_r})$

density function (Fig. 2.2), this is equivalent to:

$$P = 1 - \Pr(Q_{\rho_1} \leq q_{\min}(\rho_1), Q_{\rho_2} \leq q_{\min}(\rho_2), \dots, Q_{\rho_R} \leq q_{\min}(\rho_R)), \quad (2.48)$$

where  $q_{\min}(\rho_r)$  is the  $1 - T^{th}$  percentile of the distribution of  $q_{\rho_r}$ , corresponding to the test statistic  $Q_{\rho_r}$ .



**Fig. 2.2** Visualisation of equivalence of Eq. 2.47 and Eq. 2.48 using the probability density function of  $q_{\rho_r}$  | Probability density function of the distribution of  $q_{\rho_r}$  (density (y-axis) as a function of  $q_{\rho_r}$  (x-axis)) highlighting the  $\Pr(Q_{\rho_r} \geq q_{\rho_r})$  which is the area under the probability density function curve to the left of  $Q_{\rho_r}$  in red (a) and additionally the greater area (as defined in Eq. 2.47) of  $1 - T$  in blue (b). The area of  $1 - T$  is bounded by  $q_{\rho_r} = q_{\min}(\rho_r)$ , where by definition  $q_{\min}(\rho_r)$  is the  $(1 - T)^{th}$  percentile of the distribution of  $q_{\rho_r}$  corresponding to the test statistic  $Q_{\rho_r}$ . It can be seen that if,  $\Pr(Q_{\rho_r} \geq q_{\rho_r}) \leq 1 - T$  it follows that  $Q_{\rho_r} \leq q_{\min}(\rho_r)$ .

I will now show that under the null,  $Q_{\rho}$ , can be written as the sum of  $S + 1$  independent random variables in the form  $\frac{1}{2}\tau_{\rho}\eta_0 + \frac{1}{2}\rho\kappa$ , where:

$$\tau_{\rho} = (1 - \rho)m + \rho \frac{\mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{Z} \mathbf{1}^T}{m}, \quad (2.49)$$

$$\mathbf{Z} = \mathbf{P}_0^{\frac{T}{2}} \text{diag}(\mathbf{x}), \quad (2.50)$$

$$m = \mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \mathbf{1}, \quad (2.51)$$

$$\kappa = \phi + \xi, \quad (2.52)$$

$$\phi = \sum_{s=1}^S \lambda_s \eta_s, \quad (2.53)$$

$$\xi = 2\mathbf{v}^T (\mathbf{I} - \mathbf{M}) \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{M} \mathbf{v}, \quad (2.54)$$

$$\mathbf{M} = \frac{1}{m} \mathbf{Z} \mathbf{1} \mathbf{1}^T \mathbf{Z}^T, \quad (2.55)$$

$$\mathbf{v} = \mathbf{P}_0^{\frac{T}{2}} \mathbf{y}, \quad (2.56)$$

$\lambda_s$  are the  $S$  non-zero eigenvalues of  $\mathbf{E}^T \mathbf{Z}^T (\mathbf{I} - \mathbf{M}) \mathbf{Z} \mathbf{E}$ , and  $\eta_s$  are i.i.d.  $\chi_1^2$  for  $s = 0, 1, \dots, S$ .

**Proof** Starting from  $Q_\rho$  (defined in Eq. 2.42):

$$Q_\rho = \frac{1}{2} \mathbf{y}^T \mathbf{P}_0 \mathbf{K}_\rho \mathbf{P}_0 \mathbf{y}, \quad (2.57)$$

substitute  $\mathbf{K}_\rho$  (as defined in Eq. 2.29) and write  $\mathbf{P}_0$  as a product of its matrix square roots to give:

$$Q_\rho = \frac{1}{2} \mathbf{y}^T \mathbf{P}_0^{\frac{1}{2}} \mathbf{P}_0^{\frac{T}{2}} [(1 - \rho) \text{diag}(\mathbf{x}) \mathbf{1} \mathbf{1}^T \text{diag}(\mathbf{x}) + \rho \text{diag}(\mathbf{x}) \mathbf{\Sigma} \text{diag}(\mathbf{x})] \mathbf{P}_0^{\frac{1}{2}} \mathbf{P}_0^{\frac{T}{2}} \mathbf{y}. \quad (2.58)$$

Let  $\mathbf{v} = \mathbf{P}_0^{\frac{T}{2}} \mathbf{y}$ , to give:

$$Q_\rho = \frac{1}{2} (1 - \rho) \mathbf{v}^T \mathbf{P}_0^{\frac{T}{2}} \text{diag}(\mathbf{x}) \mathbf{1} \mathbf{1}^T \text{diag}(\mathbf{x}) \mathbf{P}_0^{\frac{1}{2}} \mathbf{v} + \frac{1}{2} \rho \mathbf{v}^T \mathbf{P}_0^{\frac{T}{2}} \text{diag}(\mathbf{x}) \mathbf{\Sigma} \text{diag}(\mathbf{x}) \mathbf{P}_0^{\frac{1}{2}} \mathbf{v} \quad (2.59)$$

and  $\mathbf{Z} = \mathbf{P}_0^{\frac{T}{2}} \text{diag}(\mathbf{x})$ , such that:

$$Q_\rho = \frac{1}{2} (1 - \rho) \mathbf{v}^T \mathbf{Z} \mathbf{1} \mathbf{1}^T \mathbf{Z}^T \mathbf{v} + \frac{1}{2} \rho \mathbf{v}^T \mathbf{Z} \mathbf{\Sigma} \mathbf{Z}^T \mathbf{v} \quad (2.60)$$

and finally, let  $\mathbf{M} = \frac{1}{m} \mathbf{Z} \mathbf{1} \mathbf{1}^T \mathbf{Z}^T$ , resulting in:

$$Q_\rho = \frac{1}{2} m (1 - \rho) \mathbf{v}^T \mathbf{M} \mathbf{v} + \frac{1}{2} \rho \mathbf{v}^T \mathbf{Z} \mathbf{\Sigma} \mathbf{Z}^T \mathbf{v}. \quad (2.61)$$

This is equivalent to:

$$\begin{aligned} Q_\rho = & \frac{1}{2} \underbrace{m(1 - \rho) \mathbf{v}^T \mathbf{M} \mathbf{v}}_{\text{Term 1}} + \frac{1}{2} \underbrace{\rho \mathbf{v}^T \mathbf{Z} \mathbf{\Sigma} \mathbf{Z}^T \mathbf{v}}_{\text{Term 2}} + \frac{1}{2} \underbrace{\rho \mathbf{v}^T \mathbf{Z} \mathbf{\Sigma} \mathbf{Z}^T \mathbf{M} \mathbf{v}}_{\text{Term 3}} - \frac{1}{2} \underbrace{\rho \mathbf{v}^T \mathbf{M} \mathbf{Z} \mathbf{\Sigma} \mathbf{Z}^T \mathbf{v}}_{\text{Term 4}} \\ & - \frac{1}{2} \underbrace{\rho \mathbf{v}^T \mathbf{M} \mathbf{Z} \mathbf{\Sigma} \mathbf{Z}^T \mathbf{M} \mathbf{v}}_{\text{Term 5}} + \frac{1}{2} \underbrace{\rho \frac{\mathbf{v}^T \mathbf{M} \mathbf{v}}{m} \times \mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \mathbf{\Sigma} \mathbf{Z}^T \mathbf{Z} \mathbf{1}}_{\text{Term 6}}, \end{aligned} \quad (2.62)$$

since Terms 3 and 4 and Terms 5 and 6 cancel out. This can be explicitly shown using the definition of  $\mathbf{M}$  (Eq. 2.55) and noting that a, b and c are scalars as

follows:

$$\begin{aligned}
\text{Term 3} &= \frac{\rho}{m} \underbrace{\mathbf{v}^T \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{Z} \mathbf{1}}_{\text{a}} \underbrace{\mathbf{1}^T \mathbf{Z}^T \mathbf{v}}_{\text{b}} \\
&= \frac{\rho}{m} \underbrace{\mathbf{v}^T \mathbf{Z} \mathbf{1}}_{\text{b}} \underbrace{\mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{v}}_{\text{a}} \\
&= \text{Term 4},
\end{aligned} \tag{2.63}$$

$$\begin{aligned}
\text{Term 5} &= \frac{\rho}{m^2} \underbrace{\mathbf{v}^T \mathbf{Z} \mathbf{1}}_{\text{b}} \underbrace{\mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{Z} \mathbf{1}}_{\text{c}} \underbrace{\mathbf{1}^T \mathbf{Z}^T \mathbf{v}}_{\text{b}} \\
&= \frac{\rho}{m^2} \underbrace{\mathbf{v}^T \mathbf{Z} \mathbf{1}}_{\text{b}} \underbrace{\mathbf{1}^T \mathbf{Z}^T \mathbf{v}}_{\text{b}} \underbrace{\mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{Z} \mathbf{1}}_{\text{c}} \\
&= \text{Term 6}.
\end{aligned} \tag{2.64}$$

Eq. 2.62 can be rewritten as:

$$\begin{aligned}
Q_\rho &= \frac{1}{2} m (1 - \rho) \mathbf{v}^T \mathbf{M} \mathbf{v} + \frac{1}{2} \rho \mathbf{v}^T \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{v} - \frac{1}{2} \rho \mathbf{v}^T \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{M} \mathbf{v} + \rho \mathbf{v}^T \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{M} \mathbf{v} \\
&\quad - \frac{1}{2} \rho \mathbf{v}^T \mathbf{M} \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{v} + \frac{1}{2} \rho \mathbf{v}^T \mathbf{M} \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{M} \mathbf{v} - \rho \mathbf{v}^T \mathbf{M} \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{M} \mathbf{v} \\
&\quad + \frac{1}{2} \rho \frac{\mathbf{v}^T \mathbf{M} \mathbf{v}}{m} \times \mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{Z} \mathbf{1},
\end{aligned} \tag{2.65}$$

which can be factorised as follows:

$$\begin{aligned}
Q_\rho &= \frac{1}{2} [m(1 - \rho) + \frac{\rho}{m} \mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{Z} \mathbf{1}] \mathbf{v}^T \mathbf{M} \mathbf{v} \\
&\quad + \frac{1}{2} \rho \mathbf{v}^T (\mathbf{I} - \mathbf{M}) \mathbf{Z} \Sigma \mathbf{Z}^T (\mathbf{I} - \mathbf{M}) \mathbf{v} + \rho \mathbf{v}^T (\mathbf{I} - \mathbf{M}) \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{M} \mathbf{v}.
\end{aligned} \tag{2.66}$$

Let  $\tau_\rho = (1 - \rho)m + \rho \frac{\mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{Z} \mathbf{1}}{m}$  and  $\xi = 2 \mathbf{v}^T (\mathbf{I} - \mathbf{M}) \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{M} \mathbf{v}$ , then:

$$Q_\rho = \frac{1}{2} \tau_\rho \underbrace{\mathbf{v}^T \mathbf{M} \mathbf{v}}_{\text{Term 1}} + \frac{1}{2} \rho \underbrace{(\mathbf{v}^T (\mathbf{I} - \mathbf{M}) \mathbf{Z} \Sigma \mathbf{Z}^T (\mathbf{I} - \mathbf{M}) \mathbf{v})}_{\text{Term 2}} + \underbrace{\xi}_{\text{Term 3}}. \tag{2.67}$$

I will now show that the three terms (noting that  $\tau_\rho$  and  $\rho$  are scalars) are independent of one another, such that  $Q_\rho$  can be expressed as a mixture of independent random variables.

Throughout these derivations, I will use the fact that  $\mathbf{v}$  is normally distributed (see Lippert *et al.*<sup>253</sup> for the proof):

$$\mathbf{v} = \mathbf{P}_0^{\frac{T}{2}} \mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2.68}$$

and let  $m = \mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \mathbf{1}$ , such that  $\mathbf{M}$  is a projection matrix (with the column space of  $\mathbf{M}$  spanned by  $\mathbf{Z}\mathbf{1}$ ).

To demonstrate the independence of the three terms, I will show that the covariance between each pair of terms is 0. Let  $\mathbf{v}$  be normally distributed (as defined in Eq. 2.68) and  $\mathbf{A}$  and  $\mathbf{B}$  be two general matrices, then using results in section 8.2 of the matrix cookbook<sup>262</sup>:

$$\begin{aligned} \text{Cov}(\mathbf{v}^T \mathbf{A} \mathbf{v}, \mathbf{v}^T \mathbf{B} \mathbf{v}) &= \mathbb{E}(\mathbf{v}^T \mathbf{A} \mathbf{v} \mathbf{v}^T \mathbf{B} \mathbf{v}) - \mathbb{E}(\mathbf{v}^T \mathbf{A} \mathbf{v}) \mathbb{E}(\mathbf{v}^T \mathbf{B} \mathbf{v}) \\ &= \text{tr}(\mathbf{A}(\mathbf{B} + \mathbf{B}^T)) + \text{tr}(\mathbf{A})\text{tr}(\mathbf{B}) - \text{tr}(\mathbf{A})\text{tr}(\mathbf{B}) \\ &= \text{tr}(\mathbf{A}(\mathbf{B} + \mathbf{B}^T)) \end{aligned} \quad (2.69)$$

Using Eq. 2.69 and noting that  $\mathbf{M}(\mathbf{I} - \mathbf{M}) = \mathbf{0}$ ,  $(\mathbf{I} - \mathbf{M}) = (\mathbf{I} - \mathbf{M})^T$  and  $\mathbf{M}\mathbf{M} = \mathbf{M}$  when  $\mathbf{M}$  is a projection matrix and  $\text{tr}(\mathbf{A}\mathbf{B}) = \text{tr}(\mathbf{B}\mathbf{A})$ :

$$\begin{aligned} \text{Cov}(\text{Term 1}, \text{Term 2}) &= \text{Cov}(\mathbf{v}^T \mathbf{M} \mathbf{v}, \mathbf{v}^T (\mathbf{I} - \mathbf{M}) \mathbf{Z} \Sigma \mathbf{Z}^T (\mathbf{I} - \mathbf{M}) \mathbf{v}) \\ &= \text{tr}(\mathbf{M}[(\mathbf{I} - \mathbf{M}) \mathbf{Z} \Sigma \mathbf{Z}^T (\mathbf{I} - \mathbf{M}) \\ &\quad + (\mathbf{I} - \mathbf{M}) \mathbf{Z} \Sigma \mathbf{Z}^T (\mathbf{I} - \mathbf{M})]) \\ &= 2\text{tr}(\mathbf{M}(\mathbf{I} - \mathbf{M}) \mathbf{Z} \Sigma \mathbf{Z}^T (\mathbf{I} - \mathbf{M})) \\ &= 0, \end{aligned} \quad (2.70)$$

$$\begin{aligned} \text{Cov}(\text{Term 1}, \text{Term 3}) &= \text{Cov}(\mathbf{v}^T \mathbf{M} \mathbf{v}, 2\mathbf{v}^T (\mathbf{I} - \mathbf{M}) \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{M} \mathbf{v}) \\ &= 2\text{tr}(\mathbf{M}[(\mathbf{I} - \mathbf{M}) \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{M} + ((\mathbf{I} - \mathbf{M}) \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{M})^T]) \\ &= 2\text{tr}(\mathbf{0} + \mathbf{M} \mathbf{M} \mathbf{Z} \Sigma \mathbf{Z}^T (\mathbf{I} - \mathbf{M})) \\ &= 2\text{tr}(\underbrace{\mathbf{M} \mathbf{Z} \Sigma \mathbf{Z}^T}_{\mathbf{A}} \underbrace{(\mathbf{I} - \mathbf{M})}_{\mathbf{B}}) \\ &= 2\text{tr}(\underbrace{(\mathbf{I} - \mathbf{M})}_{\mathbf{B}} \underbrace{\mathbf{M} \mathbf{Z} \Sigma \mathbf{Z}^T}_{\mathbf{A}}) \\ &= 0 \end{aligned} \quad (2.71)$$

and

$$\begin{aligned}
& \text{Cov}(\text{Term 2}, \text{Term 3}) \\
&= \text{Cov}(\mathbf{v}^T(\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T(\mathbf{I} - \mathbf{M})\mathbf{v}, 2\mathbf{v}^T(\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T\mathbf{M}\mathbf{v}) \\
&= 2\text{tr}((\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T(\mathbf{I} - \mathbf{M})[(\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T\mathbf{M} + ((\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T\mathbf{M})^T]) \\
&= 2\text{tr}((\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T(\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T\mathbf{M} \\
&\quad + (\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T(\mathbf{I} - \mathbf{M})\mathbf{M}\mathbf{Z}\Sigma\mathbf{Z}^T(\mathbf{I} - \mathbf{M})) \\
&= 2\text{tr}(\underbrace{(\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T(\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T}_{\mathbf{A}}\underbrace{\mathbf{M}}_{\mathbf{B}} + 0) \\
&= 2\text{tr}(\underbrace{\mathbf{M}}_{\mathbf{B}}\underbrace{(\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T(\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T}_{\mathbf{A}}) \\
&= 0.
\end{aligned} \tag{2.72}$$

Using that when  $\mathbf{A}$  is a symmetric matrix then:

$$\mathbf{v}^T\mathbf{A}\mathbf{v} \sim \sum_k \lambda_k \chi_1^2, \tag{2.73}$$

where  $\lambda_k$  are the eigenvalues of matrix  $\mathbf{A}$  (see Lippert *et al.*<sup>253</sup> for the proof), I now show that the first two terms of Eq. 2.67 follow mixtures of  $\chi_1^2$  variables.

Since the only eigenvalues of a projection matrix  $\mathbf{M}$  are 0 and 1, then Term 1,  $\mathbf{v}^T\mathbf{M}\mathbf{v}$ , has the following distribution:

$$\mathbf{v}^T\mathbf{M}\mathbf{v} \sim \chi_1^2, \tag{2.74}$$

Term 2,  $\mathbf{v}^T(\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T(\mathbf{I} - \mathbf{M})\mathbf{v}$ , follows the distribution:

$$\mathbf{v}^T(\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T(\mathbf{I} - \mathbf{M})\mathbf{v} \sim \sum_k \lambda_k \chi_1^2, \tag{2.75}$$

where  $\lambda_k$  are the eigenvalues of  $\mathbf{E}^T\mathbf{Z}^T(\mathbf{I} - \mathbf{M})\mathbf{Z}\mathbf{E}$ , where I have used that

eigenvalues( $\mathbf{A}\mathbf{A}^T$ ) = eigenvalues( $\mathbf{A}^T\mathbf{A}$ )<sup>†</sup>, such that:

$$\begin{aligned}
& \text{eigenvalues}((\mathbf{I} - \mathbf{M})\mathbf{Z}\mathbf{\Sigma}\mathbf{Z}^T(\mathbf{I} - \mathbf{M})) \\
&= \text{eigenvalues}(\underbrace{(\mathbf{I} - \mathbf{M})\mathbf{Z}\mathbf{E}}_{\mathbf{A}}\underbrace{\mathbf{E}^T\mathbf{Z}^T(\mathbf{I} - \mathbf{M})}_{\mathbf{A}^T}) \\
&= \text{eigenvalues}(\underbrace{\mathbf{E}^T\mathbf{Z}^T(\mathbf{I} - \mathbf{M})}_{\mathbf{A}^T}\underbrace{(\mathbf{I} - \mathbf{M})\mathbf{Z}\mathbf{E}}_{\mathbf{A}}) \\
&= \text{eigenvalues}(\mathbf{E}^T\mathbf{Z}^T(\mathbf{I} - \mathbf{M})\mathbf{Z}\mathbf{E}). \tag{2.76}
\end{aligned}$$

Therefore, I can write Eq. 2.67 as:

$$Q_\rho = \frac{1}{2}\tau_\rho\eta_0 + \frac{1}{2}\rho\kappa, \tag{2.77}$$

where  $\kappa = \phi + \xi$  and  $\phi = \sum_{s=1}^S \lambda_s \eta_s$ , where  $\lambda_s$  are the non-zero eigenvalues of  $\mathbf{E}^T\mathbf{Z}^T(\mathbf{I} - \mathbf{M})\mathbf{Z}\mathbf{E}$ , and  $\eta_s$  are i.i.d.  $\chi_1^2$  for  $s = 0, 1, \dots, S$ .

Having proved that  $Q_\rho = \frac{1}{2}\tau_\rho\eta_0 + \frac{1}{2}\rho\kappa$ , I now substitute this form for  $Q_\rho$  into Eq. 2.48, to give:

$$P = 1 - \Pr\left(\left(\frac{1}{2}\tau_{\rho_1}\eta_0 + \frac{1}{2}\rho_1\kappa\right) \leq q_{\min}(\rho_1), \dots, \left(\frac{1}{2}\tau_{\rho_R}\eta_0 + \frac{1}{2}\rho_R\kappa\right) \leq q_{\min}(\rho_R)\right), \tag{2.78}$$

which can be rewritten in terms of  $\kappa$ , as:

$$\begin{aligned}
P &= 1 - \mathbb{E}[\Pr(\kappa \leq \frac{(2q_{\min}(\rho_1) - \tau_{\rho_1}\eta_0)}{\rho_1}, \dots, \kappa \leq \frac{(2q_{\min}(\rho_R) - \tau_{\rho_R}\eta_0)}{\rho_R}) | \eta_0)] \\
&= 1 - \mathbb{E}[\Pr(\kappa \leq \min_{r=1}^R \left\{ \frac{(2q_{\min}(\rho_r) - \tau_{\rho_r}\eta_0)}{\rho_r} \right\}) | \eta_0] \tag{2.79}
\end{aligned}$$

Again using results from section 8.2 of the matrix cookbook<sup>262</sup>,  $tr(\mathbf{A}\mathbf{B}) = tr(\mathbf{B}\mathbf{A})$ ,  $\mathbf{M}(\mathbf{I} - \mathbf{M}) = \mathbf{0}$  and  $(\mathbf{I} - \mathbf{M})(\mathbf{I} - \mathbf{M}) = (\mathbf{I} - \mathbf{M})$  ( $\mathbf{M}$  is a projection matrix), it can be shown that:

$$\mathbb{E}(\xi) = \mathbb{E}(2\mathbf{v}^T(\mathbf{I} - \mathbf{M})\mathbf{Z}\mathbf{\Sigma}\mathbf{Z}^T\mathbf{M}\mathbf{v}) = 2tr((\mathbf{I} - \mathbf{M})\mathbf{Z}\mathbf{\Sigma}\mathbf{Z}^T\mathbf{M}) = 0 \tag{2.80}$$

---

<sup>†</sup>and noting that  $\mathbf{\Sigma} = \mathbf{E}\mathbf{E}^T$

and

$$\begin{aligned}
\text{Var}(\xi) &= \text{Var}(2\mathbf{v}^T(\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T\mathbf{M}\mathbf{v}) \\
&= 4\text{tr}([( \mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T\mathbf{M}][(\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T\mathbf{M} + ((\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T\mathbf{M})^T]) \\
&= 4\text{tr}((\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T\mathbf{M}(\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T\mathbf{M} \\
&\quad + (\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T\mathbf{M}\mathbf{M}^T\mathbf{Z}\Sigma\mathbf{Z}^T(\mathbf{I} - \mathbf{M})^T) \\
&= 4\text{tr}(0 + (\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T\mathbf{M}\mathbf{Z}\Sigma\mathbf{Z}^T(\mathbf{I} - \mathbf{M})) \\
&= 4\text{tr}((\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T\mathbf{M}\mathbf{Z}\Sigma\mathbf{Z}^T(\mathbf{I} - \mathbf{M})) \\
&= 4\text{tr}((\mathbf{I} - \mathbf{M})\mathbf{Z}\Sigma\mathbf{Z}^T\mathbf{M}\mathbf{Z}\Sigma\mathbf{Z}^T) \\
&= 4\text{tr}((\mathbf{I} - \mathbf{M})\mathbf{Z}\mathbf{E}\mathbf{E}^T\mathbf{Z}^T\mathbf{M}\mathbf{Z}\mathbf{E}\mathbf{E}^T\mathbf{Z}^T) \\
&= 4\text{tr}(\mathbf{E}^T\mathbf{Z}^T(\mathbf{I} - \mathbf{M})\mathbf{Z}\mathbf{E}\mathbf{E}^T\mathbf{Z}^T\mathbf{M}\mathbf{Z}\mathbf{E}) \tag{2.81}
\end{aligned}$$

It therefore follows that the P value (as defined by Eq. 2.79) can be calculated using one-dimensional numerical integration, approximating the distribution of  $\kappa$  using Davies exact characteristic inversion method<sup>123,260</sup> after adjusting for the extra variance term of  $\xi$ , switching to the modified Liu method<sup>123,258</sup> when Davies method fails to converge.

## 2.2.6 Computational complexities

As described in Section 2.2.5, the StructLMM interaction and joint association tests require fitting null models, calculating test statistics and then computing the corresponding P values. These operations are implemented in a computationally efficient manner, such that the tests scale linearly with the number of samples (assuming that  $N \gg L > a$  where  $N$  is the number of samples,  $L$  the number of environments and  $a$  the number of covariates) instead of cubically as would be the case if implemented naively.

In this section, I will first describe the steps and computational efficiency of general operations that are used repeatedly, followed by computational details for each of the tasks required for the StructLMM tests. I then provide a table summarising the computational complexity of the main steps involved.

### Computational complexity of general operations

I will first describe how  $\mathbf{P}_0\mathbf{A}$  is computed efficiently, where  $\mathbf{P}_0$  is as defined in Eq. 2.37 and  $\mathbf{A}$  is a general  $N \times M$  matrix (or an  $N \times 1$  vector, corresponding to



the special case that  $M = 1$ ), under the assumption that  $M \ll N$ .

Explicitly,

$$\mathbf{P}_0 \mathbf{A} = \underbrace{\mathbf{K}_0^{-1} \mathbf{A}}_{\text{Term 1}} - \underbrace{\mathbf{K}_0^{-1} \mathbf{X} (\mathbf{X}^T \mathbf{K}_0^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{K}_0^{-1} \mathbf{A}}_{\text{Term 2}}. \quad (2.82)$$

I will start with the calculation of Term 1. Whilst naive calculation of  $\mathbf{K}_0^{-1} \mathbf{A}$  is cubic in the number of samples,  $N$ , since  $\mathbf{K}_0$  (defined in Eq. 2.35) is the combination of a low rank matrix (rank  $L$ ) and an identity matrix, the computational complexity of this operation can be reduced<sup>76,78,253</sup>, by writing  $\mathbf{K}_0$  as follows:

Let  $\delta = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_n^2}$ . Then:

$$\mathbf{K}_0^{-1} = (\hat{\sigma}_n^2 (\delta \mathbf{E} \mathbf{E}^T + \mathbf{I}_N))^{-1}. \quad (2.83)$$

Let the spectral decomposition of  $\mathbf{E} \mathbf{E}^T = \mathbf{U} \mathbf{S} \mathbf{U}^T$ , noting that  $\mathbf{U} \mathbf{U}^T = \mathbf{I}_N$ , then:

$$\mathbf{K}_0^{-1} = (\hat{\sigma}_n^2 \mathbf{U} (\delta \mathbf{S} + \mathbf{I}_N) \mathbf{U}^T)^{-1}. \quad (2.84)$$

Using  $(\mathbf{A} \mathbf{B})^{-1} = \mathbf{B}^{-1} \mathbf{A}^{-1}$ ,  $\mathbf{U}^{-1} = \mathbf{U}^T$  and  $\mathbf{U}^{-T} = \mathbf{U}$ :

$$\mathbf{K}_0^{-1} = \frac{1}{\hat{\sigma}_n^2} \mathbf{U} (\delta \mathbf{S} + \mathbf{I}_N)^{-1} \mathbf{U}^T. \quad (2.85)$$

Since  $\mathbf{E} \mathbf{E}^T$  is a low rank matrix, with rank  $L$ :

$$\mathbf{U} (\delta \mathbf{S} + \mathbf{I}_N)^{-1} \mathbf{U}^T = [\mathbf{U}_1, \mathbf{U}_2] \begin{bmatrix} \delta \mathbf{S} + \mathbf{I}_L & 0 \\ 0 & \mathbf{I}_{N-L} \end{bmatrix}^{-1} [\mathbf{U}_1, \mathbf{U}_2]^T, \quad (2.86)$$

where  $\mathbf{U}_1$  is an  $N \times L$  matrix and  $\mathbf{U}_2$  is an  $N \times (N - L)$  matrix. As the inversion operation is of a diagonal matrix, this can be rewritten as:

$$\mathbf{U}_1 (\delta \mathbf{S} + \mathbf{I}_L)^{-1} \mathbf{U}_1^T + \mathbf{U}_2 \mathbf{I}_{N-L} \mathbf{U}_2^T = \mathbf{U}_1 (\delta \mathbf{S} + \mathbf{I}_L)^{-1} \mathbf{U}_1^T + \mathbf{U}_2 \mathbf{U}_2^T. \quad (2.87)$$

Additionally as  $\mathbf{U} \mathbf{U}^T = \mathbf{I}_N$ , this implies that:

$$\mathbf{U} \mathbf{U}^T = [\mathbf{U}_1, \mathbf{U}_2] [\mathbf{U}_1, \mathbf{U}_2]^T = \mathbf{U}_1 \mathbf{U}_1^T + \mathbf{U}_2 \mathbf{U}_2^T = \mathbf{I}_N \quad (2.88)$$

and therefore:

$$\mathbf{U}_2 \mathbf{U}_2^T = \mathbf{I}_N - \mathbf{U}_1 \mathbf{U}_1^T, \quad (2.89)$$

resulting in:

$$\begin{aligned}\mathbf{K}_0^{-1} &= \frac{1}{\hat{\sigma}_n^2} [\mathbf{U}_1(\delta\mathbf{S} + \mathbf{I}_L)^{-1}\mathbf{U}_1^T + (\mathbf{I}_N - \mathbf{U}_1\mathbf{U}_1^T)(\mathbf{I}_N - \mathbf{U}_1\mathbf{U}_1^T)^T] \\ &= \frac{1}{\hat{\sigma}_n^2} [\mathbf{U}_1(\delta\mathbf{S} + \mathbf{I}_L)^{-1}\mathbf{U}_1^T + (\mathbf{I}_N - \mathbf{U}_1\mathbf{U}_1^T)(\mathbf{I}_N - \mathbf{U}_1\mathbf{U}_1^T)].\end{aligned}\quad (2.90)$$

Using the form of  $\mathbf{K}_0^{-1}$  given by Eq. 2.90, calculation of Term 1,  $\mathbf{K}_0^{-1}\mathbf{A}$ , can be written as:

$$\begin{aligned}\mathbf{K}_0^{-1}\mathbf{A} &= \frac{1}{\hat{\sigma}_n^2} [\mathbf{U}_1(\delta\mathbf{S} + \mathbf{I}_L)^{-1}\mathbf{U}_1^T + (\mathbf{I}_N - \mathbf{U}_1\mathbf{U}_1^T)(\mathbf{I}_N - \mathbf{U}_1\mathbf{U}_1^T)]\mathbf{A} \\ &= \frac{1}{\hat{\sigma}_n^2} [\mathbf{U}_1(\delta\mathbf{S} + \mathbf{I}_L)^{-1}\mathbf{U}_1^T\mathbf{A} + (\mathbf{I}_N - \mathbf{U}_1\mathbf{U}_1^T)(\mathbf{I}_N - \mathbf{U}_1\mathbf{U}_1^T)\mathbf{A}] \\ &= \frac{1}{\hat{\sigma}_n^2} [\underbrace{\mathbf{U}_1(\delta\mathbf{S} + \mathbf{I}_L)^{-1}\mathbf{U}_1^T\mathbf{A}}_{\text{Term a}} + \underbrace{(\mathbf{I}_N - \mathbf{U}_1\mathbf{U}_1^T)(\mathbf{A} - \mathbf{U}_1\mathbf{U}_1^T\mathbf{A})}_{\text{Term b}}].\end{aligned}\quad (2.91)$$

Spectral decomposition of the low rank matrix  $\mathbf{E}\mathbf{E}^T$  to find the eigenvectors  $\mathbf{U}_1$  and eigenvalues  $\mathbf{S}$  is  $O(NL^2)^{263}$  and inversion of the  $L \times L$  matrix,  $(\delta\mathbf{S} + \mathbf{I}_L)$  is  $O(L)$ . Calculation of  $\mathbf{U}_1^T\mathbf{A}$  and  $\mathbf{U}_1(\delta\mathbf{S} + \mathbf{I}_L)^{-1}$  are  $O(NLM)$  and  $O(NL^2)$ , respectively and then matrix multiplication of the two aforementioned terms gives, Term a,  $\mathbf{U}_1(\delta\mathbf{S} + \mathbf{I}_L)^{-1}\mathbf{U}_1^T\mathbf{A}$  with  $O(NLM)$ . Having already calculated,  $\mathbf{U}_1^T\mathbf{A}$ , matrix multiplication with  $\mathbf{U}_1$  to give  $\mathbf{U}_1\mathbf{U}_1^T\mathbf{A}$  is  $O(NLM)$ . This is followed by matrix subtraction of  $O(NM)$  such that  $(\mathbf{A} - \mathbf{U}_1\mathbf{U}_1^T\mathbf{A})$  is a  $N \times M$  matrix and thus the complexity of calculating Term b,  $(\mathbf{I}_N - \mathbf{U}_1\mathbf{U}_1^T)(\mathbf{A} - \mathbf{U}_1\mathbf{U}_1^T\mathbf{A})$  (once the second bracketed term has been computed) is the same as the complexity of computing  $(\mathbf{I}_N - \mathbf{U}_1\mathbf{U}_1^T)\mathbf{A}$ , which has identical form to the expression in the second bracket of Term b, for which I have already described the computational complexity. The final operation to calculate  $\mathbf{K}_0^{-1}\mathbf{A}$  is the addition of two  $N \times M$  matrices which has  $O(NM)$ . Therefore the complexity of calculating  $\mathbf{K}_0^{-1}\mathbf{A}$  has  $O(NL^2 + NLM)$ .

I will describe the computational complexity to calculate Term 2 in Eq. 2.82. First, recall that  $\mathbf{X}$  is the design matrix of the covariates which is  $N \times (a + 1)$  for the interaction test ( $\mathbf{X} = [\mathbf{W}, \mathbf{x}]$  as described in Section 2.2.5) and  $N \times a$  for the joint association test ( $\mathbf{X} = \mathbf{W}$  as described in Section 2.2.5). For simplicity, the complexity of any operations involving  $\mathbf{X}$ , will be described on the basis that  $\mathbf{X}$  is an  $N \times a$  matrix. I will now split Term 2 as follows:

$$\text{Term 2} = \underbrace{\mathbf{K}_0^{-1}\mathbf{X}}_{\text{Term a}} \underbrace{(\mathbf{X}^T \mathbf{K}_0^{-1} \mathbf{X})^{-1}}_{\text{Term a}} \underbrace{\mathbf{X}^T \mathbf{K}_0^{-1} \mathbf{A}}_{\text{Term b}}. \quad (2.92)$$

Terms a and b, have the same form as Term 1 (previously described) and hence have computational complexity  $O(NL^2 + NLa)$  and  $O(NL^2 + NLM)$ , respectively. Calculation of Term c, requires matrix multiplication of  $\mathbf{X}^T$  with Term a which is  $O(Na^2)$  followed by inversion of an  $a \times a$  matrix which is  $O(a^3)$ . Calculation of Term d, requires matrix multiplication of  $\mathbf{X}^T$  with Term b which is  $O(NMa)$ . Matrix multiplication of Term a and Term c is  $O(Na^2)$  and then further matrix multiplication with Term d is  $O(NMa)$ . Thus calculation of Term 2 is  $O(NL^2 + NLM + NLa + NMa + Na^2 + a^3)$ .

Therefore calculation of  $\mathbf{P}_0\mathbf{A}$ , where  $\mathbf{A}$  is a general  $N \times M$  matrix has computational complexity  $O(NL^2 + NLM + NLa + NMa + Na^2 + a^3)$ .

From this it follows that calculation of  $\mathbf{BP}_0\mathbf{A}$ , where again  $\mathbf{P}_0$  is as defined in Eq. 2.37 and  $\mathbf{A}$  is a general  $N \times M$  matrix and  $\mathbf{B}$  is another general  $P \times N$  matrix (also assuming that  $P \ll N$ ), involves a further matrix multiplication operation of  $O(NMP)$  such that the total computational complexity to compute  $\mathbf{BP}_0\mathbf{A}$  is  $O(NL^2 + NLM + NLa + NMa + Na^2 + a^3 + NMP)$ .

### Computational complexity of fitting the null model

The StructLMM interaction test and joint association test null models (Eq. 2.32 and Eq. 2.34, respectively) are fitted using LIMIX (using REML-based inference)<sup>78</sup>, which is implemented efficiently. In particular the linear covariance structure of  $\mathbf{\Sigma}$ , such that the covariance of the null distribution for both the interaction and joint association test can be written as  $\sigma_e^2\mathbf{E}\mathbf{E}^T + \sigma_n^2\mathbf{I}_N$  (see Eqs. 2.32 and 2.34), is exploited enabling fitting of these null models with computational complexity  $O(NL^2 + L^3)$  (I refer the reader to Lippert *et al.*<sup>253</sup> and Casale *et al.*<sup>79</sup> for full details).

I note that for the interaction test, a null model per variant needs to be fitted (see Eq. 2.32), whilst for the association test only one null model for all considered variants is required.

## Computational complexity of calculating $Q_\rho$

The score-based test statistic  $Q_\rho$  (Eq. 2.42), with the form of  $\mathbf{K}_\rho$  described by Eq. 2.30 explicitly substituted gives:

$$\begin{aligned} Q_\rho &= \frac{1}{2} \mathbf{y}^T \mathbf{P}_0 [(1 - \rho) \mathbf{x} \mathbf{x}^T + \rho \text{diag}(\mathbf{x}) \boldsymbol{\Sigma} \text{diag}(\mathbf{x})] \mathbf{P}_0 \mathbf{y} \\ &= \left( \frac{1}{\sqrt{2}} \mathbf{y}^T \mathbf{P}_0 \mathbf{D} \right) \left( \frac{1}{\sqrt{2}} \mathbf{y}^T \mathbf{P}_0 \mathbf{D} \right)^T, \end{aligned} \quad (2.93)$$

where  $\mathbf{D} = [\sqrt{1 - \rho} \mathbf{x}, \sqrt{\rho} \text{diag}(\mathbf{x}) \mathbf{E}]$ , noting that the interaction test statistics has the same form with  $\mathbf{K}_\rho = \mathbf{K}_1$ , such that  $\mathbf{D}$  is an  $N \times L$  matrix for the interaction test and  $N \times (L + 1)$  for the association test. For simplicity, I will use the dimensions of the former when stating the computational complexities.

Using  $\text{diag}(\mathbf{x}) \mathbf{E} = \mathbf{x} \odot \mathbf{E}$ , this operation has complexity  $O(NL)$ . Then noting that  $\mathbf{y}^T \mathbf{P}_0 \mathbf{D}$  has the same form as  $\mathbf{B} \mathbf{P}_0 \mathbf{A}$  (general computation described above) with  $P = 1$  and  $M = L$ , this can be calculated in  $O(NL^2 + NLa + Na^2 + a^3)$ . Then matrix multiplication of  $\mathbf{B} \mathbf{P}_0 \mathbf{A}$  with its transpose is  $O(L^2)$ .

Thus computation of this score-based test statistic has complexity  $O(NL^2 + NLa + Na^2 + a^3)$ .

## Computational complexity of evaluating the significance of the score-based test statistics

The eigenvalues of  $\frac{1}{2} \mathbf{P}_0^{\frac{T}{2}} \mathbf{D} \mathbf{D}^T \mathbf{P}_0^{\frac{1}{2}}$  are required to calculate the P values of the score-based test statistics (see Section 2.2.5). Naively, this requires eigenvalue decomposition of an  $N \times N$  matrix which has computational complexity of  $O(N^3)$ .

However, using the fact that  $\text{eigenvalues}(\mathbf{A} \mathbf{A}^T) = \text{eigenvalues}(\mathbf{A}^T \mathbf{A})$ <sup>262</sup>, I instead calculate the eigenvalues of,  $\frac{1}{2} \mathbf{D}^T \mathbf{P}_0^{\frac{1}{2}} \mathbf{P}_0^{\frac{T}{2}} \mathbf{D} = \frac{1}{2} \mathbf{D}^T \mathbf{P}_0 \mathbf{D}$ . Noting that  $\mathbf{D}^T \mathbf{P}_0 \mathbf{D}$  has the same form as  $\mathbf{B} \mathbf{P}_0 \mathbf{A}$  (general computation described above) with  $P = M = L$ , this can be calculated as  $O(NL^2 + NLa + Na^2 + a^3)$  and eigenvalue decomposition of the resulting  $L \times L$  matrix has computational complexity  $O(L^3)$ .

Calculation of P values using the exact Davies<sup>260</sup> or the modified moment approximation method<sup>123</sup> are independent of the sample size,  $N$ .

Therefore this operation is  $O(NL^2 + NLa + Na^2 + L^3 + a^3)$ .

## Computational complexity of the interaction test

To summarise the computational complexity of the interaction test is  $O(NL^2 + NLa + Na^2 + L^3 + a^3)$  and thus scales linearly with the number of samples,  $N$ .

## Computational complexity of the association test

A number of operations involved in calculating,  $Q_\rho$  and the corresponding P value,  $P_\rho$ , will need to be calculated  $R$  times (once for each of the  $R$  values for  $\rho$ ). Thus these initial steps are computed with complexity  $O(NL^2R + NLaR + Na^2R + a^3R)$ , noting that  $R \ll N$  (and in many cases  $R \ll L$ ).

To calculate the final P value for the association test, corresponding to the test statistic,  $T$ ,  $\tau_\rho$  and the quantities defining  $\kappa$  need to be computed.

I will first consider  $\tau_\rho = (1 - \rho)m + \rho \frac{\mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{Z} \mathbf{1}^T}{m}$ , where  $m = \mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \mathbf{1}$  and  $\mathbf{Z} = \mathbf{P}_0^{\frac{T}{2}} \text{diag}(\mathbf{x})$ .

$m$ , can be written as follows:

$$m = \mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \mathbf{1} = \mathbf{1}^T \text{diag}(\mathbf{x}) \mathbf{P}_0^{\frac{1}{2}} \mathbf{P}_0^{\frac{T}{2}} \text{diag}(\mathbf{x}) \mathbf{1} = \mathbf{x}^T \mathbf{P}_0 \mathbf{x}. \quad (2.94)$$

Noting that  $\mathbf{x}^T \mathbf{P}_0 \mathbf{x}$  has the same form as  $\mathbf{B} \mathbf{P}_0 \mathbf{A}$  (general computation described above) with  $P = M = 1$ , this can be calculated in  $O(NL^2 + NLa + Na^2 + a^3)$ .

Now consider, the numerator of the second term of  $\tau_\rho$ ,  $\mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{Z} \mathbf{1}^T$ , which can be written as follows:

$$\begin{aligned} \mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \Sigma \mathbf{Z}^T \mathbf{Z} \mathbf{1}^T &= \mathbf{1}^T \text{diag}(\mathbf{x}) \mathbf{P}_0^{\frac{1}{2}} \mathbf{P}_0^{\frac{T}{2}} \text{diag}(\mathbf{x}) \mathbf{E} \mathbf{E}^T \text{diag}(\mathbf{x}) \mathbf{P}_0^{\frac{1}{2}} \mathbf{P}_0^{\frac{T}{2}} \text{diag}(\mathbf{x}) \mathbf{1} \\ &= (\mathbf{x}^T \mathbf{P}_0 (\mathbf{x} \odot \mathbf{E})) (\mathbf{x}^T \mathbf{P}_0 (\mathbf{x} \odot \mathbf{E}))^T. \end{aligned} \quad (2.95)$$

I have already calculated  $\mathbf{x} \odot \mathbf{E}$  (in order to calculate  $Q_\rho$ ) which gives a  $N \times L$  matrix and calculation of  $\mathbf{x}^T \mathbf{P}_0 (\mathbf{x} \odot \mathbf{E})$  is analogous to calculation of  $\mathbf{B} \mathbf{P}_0 \mathbf{A}$  (general computation described above) with  $P = 1$  and  $M = L$ , which can be computed with complexity  $O(NL^2 + NLa + Na^2 + a^3)$ . Matrix multiplication of  $(\mathbf{x} \odot \mathbf{E})^T \mathbf{P}_0 \mathbf{x}$  with its transpose is  $O(L^2)$ .

Therefore  $\tau_\rho$  is computed with complexity  $O(NL^2 + NLa + Na^2 + a^3)$ .

$\kappa$  is defined as  $\kappa = \phi + \xi$ . I need to calculate the eigenvalues corresponding to the

variable  $\phi$ :

$$\begin{aligned}
& \mathbf{E}^T \mathbf{Z}^T (\mathbf{I} - \mathbf{M}) \mathbf{Z} \mathbf{E} \\
&= \mathbf{E}^T \text{diag}(\mathbf{x}) \mathbf{P}_0^{\frac{1}{2}} (\mathbf{I} - \mathbf{M}) \mathbf{P}_0^{\frac{T}{2}} \text{diag}(\mathbf{x}) \mathbf{E} \\
&= (\mathbf{x} \odot \mathbf{E})^T \mathbf{P}_0^{\frac{1}{2}} \mathbf{P}_0^{\frac{T}{2}} (\mathbf{x} \odot \mathbf{E}) - (\mathbf{x} \odot \mathbf{E})^T \mathbf{P}_0^{\frac{1}{2}} \mathbf{M} \mathbf{P}_0^{\frac{T}{2}} (\mathbf{x} \odot \mathbf{E}) \\
&= (\mathbf{x} \odot \mathbf{E})^T \mathbf{P}_0 (\mathbf{x} \odot \mathbf{E}) - \frac{1}{m} (\mathbf{x} \odot \mathbf{E})^T \mathbf{P}_0^{\frac{1}{2}} \mathbf{P}_0^{\frac{T}{2}} \text{diag}(\mathbf{x}) \mathbf{1} \mathbf{1}^T \text{diag}(\mathbf{x}) \mathbf{P}_0^{\frac{1}{2}} \mathbf{P}_0^{\frac{T}{2}} (\mathbf{x} \odot \mathbf{E}) \\
&= (\mathbf{x} \odot \mathbf{E})^T \mathbf{P}_0 (\mathbf{x} \odot \mathbf{E}) - \frac{1}{m} (\mathbf{x} \odot \mathbf{E})^T \mathbf{P}_0 \mathbf{x} \mathbf{x}^T \mathbf{P}_0 (\mathbf{x} \odot \mathbf{E}) \\
&= (\mathbf{x} \odot \mathbf{E})^T \mathbf{P}_0 (\mathbf{x} \odot \mathbf{E}) - \frac{1}{m} ((\mathbf{x} \odot \mathbf{E})^T \mathbf{P}_0 \mathbf{x}) ((\mathbf{x} \odot \mathbf{E})^T \mathbf{P}_0 \mathbf{x})^T, \tag{2.96}
\end{aligned}$$

using  $\mathbf{M}$  as defined in Eq. 2.55.  $(\mathbf{x} \odot \mathbf{E})^T \mathbf{P}_0 (\mathbf{x} \odot \mathbf{E})$  has the same form as  $\mathbf{B} \mathbf{P}_0 \mathbf{A}$  (general computation described above) with  $P = M = L$ , which can be computed with complexity  $O(NL^2 + NLa + Na^2 + a^3)$  and  $(\mathbf{x} \odot \mathbf{E})^T \mathbf{P}_0 \mathbf{x} ((\mathbf{x} \odot \mathbf{E})^T \mathbf{P}_0 \mathbf{x})^T$  has already been computed (in order to calculate  $\tau_\rho$ ). Eigenvalue decomposition of the resulting  $L \times L$  matrix is  $O(L^3)$ .

I also need to calculate:

$$\text{Var}(\xi) = 4 \text{tr} \left( \underbrace{\mathbf{E}^T \mathbf{Z}^T (\mathbf{I} - \mathbf{M}) \mathbf{Z} \mathbf{E}}_{\text{Term 1}} \underbrace{\mathbf{E}^T \mathbf{Z}^T \mathbf{M} \mathbf{Z} \mathbf{E}}_{\text{Term 2}} \right), \tag{2.97}$$

which requires matrix multiplication of Term 1, which is the term just computed to calculate the eigenvalues of  $\phi$  and Term 2 which is one of the two terms used to calculate the eigenvalues of  $\phi$  (the second term above) and therefore has complexity  $O(L^3)$ . Finally calculation of the trace is  $O(L)$ .

Therefore calculation of the terms needed to define the variable  $\kappa$  is  $O(NL^2 + NLa + Na^2 + L^3 + a^3)$ .

Calculation of P values using the exact Davies<sup>260</sup> or the modified moment approximation method<sup>123</sup> whilst adjusting for the extra variance of  $\xi$  is independent of the sample size,  $N$ .

To summarise the computational complexity of the StructLMM joint association method is  $O(NL^2R + NLaR + Na^2R + L^3R + a^3R)$  and thus scales linearly with the number of samples,  $N$ .

## Summary of the computational complexity of the StructLMM interaction and association test

Table 2.1 summarises the computational complexity of different operations required for the StructLMM interaction and association tests, based on  $N$  individuals,  $L$  environments,  $a$  covariates and for the association test,  $R$  values of  $\rho$  to be grid searched over, assuming that  $N \gg L > a$  (also assuming that  $N \gg R$  but making no assumptions on the relationship between  $R$ ,  $L$  and  $a$ ).

Operation	Computational complexity	Interaction test	Association test
Fitting null model	$O(NL^2 + L^3)$	Once per variant	Once for all variants genome-wide
Calculation of score-based test statistic, $Q_\rho$	$O(NL^2 + NLa + Na^2 + a^3)$	Once per variant	$R$ times per variant
Evaluating the significance of the score-based test statistic	$O(NL^2 + NLa + Na^2 + L^3 + a^3)$	Once per variant	$R$ times per variant
Calculation of $\tau_\rho$	$O(NL^2 + NLa + Na^2 + a^3)$	N/A	$R$ times per variant
Calculation of $\kappa$	$O(NL^2 + NLa + Na^2 + L^3 + a^3)$	N/A	Once per variant
Total complexity of StructLMM interaction test	$O(NL^2 + NLa + Na^2 + L^3 + a^3)$	Per variant	N/A
Total complexity of StructLMM association test	$O(NL^2R + NLaR + Na^2R + L^3R + a^3R)$	N/A	Per variant

**Table 2.1 Computational complexity of different operations required for the StructLMM interaction and association tests** | Summary of the computational complexity of different operations required for the StructLMM interaction and association tests, showing the operation (column 1), the order of computational complexity (column 2) and the number of times an operation needs to be repeated for the interaction test (column 3) and association test (column 4), based on  $N$  individuals,  $L$  environments,  $a$  covariates and  $R$  values of  $\rho$  to be grid searched over in the association test, assuming that  $N \gg L > a$  (also assuming that  $N \gg R$  but making no assumptions on the relationship between  $R$ ,  $L$  and  $a$ ).

### 2.2.7 Relationship to existing methods

In this section, I will place StructLMM in context, comparing to existing interaction (and joint association) tests and describe technical similarities, where appropriate.

Interaction tests between a single genetic variant and a single environmental variable are already established (see Section 2.1) and the underlying model can be cast

as:

$$\text{logit}(\mathbf{y}_D) = \mathbf{W}\boldsymbol{\alpha} + \mathbf{x}\beta_G + \mathbf{x} \odot \mathbf{e}\beta_{G \times E} + \mathbf{e}\beta_E + \boldsymbol{\epsilon}, \quad (2.98)$$

where  $\mathbf{y}_D$  is an  $N \times 1$  binary phenotype vector, capturing for example disease status<sup>179,218</sup> or:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \mathbf{x}\beta_G + \mathbf{x} \odot \mathbf{e}\beta_{G \times E} + \mathbf{e}\beta_E + \boldsymbol{\epsilon}, \quad (2.99)$$

where  $\mathbf{y}$  is an  $N \times 1$  quantitative phenotype vector. In both cases,  $N$  is the number of samples used in the analysis,  $\mathbf{W}$  is an  $N \times a$  covariate design matrix and  $\boldsymbol{\alpha}$  is an  $a \times 1$  vector of corresponding effects,  $\mathbf{x}$  is an  $N \times 1$  genotype vector for the focal variant,  $\mathbf{e}$  is an  $N \times 1$  environmental exposure vector, and  $\beta_G, \beta_{G \times E}$  and  $\beta_E$  are the marginal genetic, interaction and environment effects, respectively.  $\boldsymbol{\epsilon}$  is an  $N \times 1$  noise vector, modelled as random effect following the multivariate normal distribution:

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}_N), \quad (2.100)$$

where  $\mathbf{I}_N$  is the  $N \times N$  identity matrix<sup>179</sup>.

For both binary and quantitative phenotypes, the model can be used to perform a 1 df interaction test (i.e.  $\beta_{G \times E} \neq 0$ ) or a 2 df joint association test  $[\beta_G, \beta_{G \times E}] \neq \mathbf{0}$ <sup>179,218</sup>.

Method development has largely focussed on increasing power to detect G×E through reducing the multiple testing burden and as outlined in Section 2.1 these methods can be classed into two main categories; those that incorporate a screening strategy, to select a subset of variants to test for interactions (between a single genetic variant and a single environmental exposure)<sup>175,193,224–231</sup> and those that use a set test based approach to test for interactions between a single environmental variable and a set of  $S$  genetic variants,  $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_S\}$ .

The underlying model for the majority of these set test based approaches can be cast as:

$$\mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \sum_{s=1}^S \mathbf{x}_s \beta_{G_s} + \sum_{s=1}^S \mathbf{x}_s \odot \mathbf{e} \beta_{(G \times E)_s} + \mathbf{e} \beta_E + \boldsymbol{\epsilon}. \quad (2.101)$$

As already discussed in Sections 1.2.1 and 2.1, not only do these set tests reduce the multiple testing burden, but in addition aggregating over variants, in particular sets of rare variants can increase the power to detect associations and interactions, when the effects are driven by multiple weak signals<sup>107–111,118,123,179</sup>.

One of the first set tests for G×E analysis was a burden based approach, proposed by Chatterjee *et al.*<sup>235</sup>, which assumes that the interaction effect is proportional to the marginal genetic and environment effects, such that  $\beta_{(G \times E)_s} = \theta \beta_{G_s} \beta_E$  in Eq. 2.101.



This is a 1 df test, since  $\theta$  takes the same value across the set of  $S$  variants, a strong constraint, further adding to the assumption that interactions only exist when there are marginal genetic and environment effects. Whilst as described in the paper, this method can in theory aggregate across both a set of variants and a set of environments, in this setting the 1 df ( $\theta$ ) is even more constraining.

The set test based approach proposed by Jiao *et al.*<sup>240</sup> relaxes this constraint in a similar vein to adaptive burden tests<sup>112–117</sup>, calculating the correlation between the variant dosages and the single environmental exposure, such that variants in the set can be weighted (by 1, -1 or 0) based on whether they are likely to drive an interaction effect (and the likely direction of this effect). Extensions include, adapting the threshold that determines if a variant is likely to have an interaction effect (i. e. determining if a variant has non-zero weight), based on the set of variants under consideration<sup>241</sup> and incorporating different filtering statistics, enabling application to quantitative phenotypes<sup>242</sup> (e. g. correlation of a genotype and environment for binary phenotypes and Levene’s statistic<sup>264</sup> for quantitative traits).

Another commonly employed burden test approach requires selecting variants to include in the genetic set based on the significance of marginal association tests (between variants and the trait of interest). It is then either assumed that all of the selected variants have the same interaction effect, i. e.  $\beta_{(G \times E)_s} = \theta_{(G \times E)}$  (and often the same association effect, i. e.  $\beta_{G_s} = \theta_G$ ) in Eq. 2.101, equivalent to first building an unweighted genetic risk score (GRS) and using this score in place of  $\mathbf{x}$  in Eq. 2.99<sup>212,213,236–239</sup>; or that the selected variants have the same interaction effect after weighting by the marginal variant effect ( $\beta_{G_s}^{\text{marg}}$ ) identified in the marginal association test, i. e.  $\beta_{(G \times E)_s} = \beta_{G_s}^{\text{marg}} \theta_{(G \times E)}$  (and similarly for the association effect, i. e.  $\beta_{G_s} = \beta_{G_s}^{\text{marg}} \theta_G$ ) in Eq. 2.101, equivalent to first building an weighted genetic risk score (GRS) and using this score in place of  $\mathbf{x}$  in Eq. 2.99<sup>207</sup>.

Variance component set tests are an alternative that can be more powerful than burden based approaches when the magnitude and direction of the interaction effects vary across the variants included within the set. As detailed below, these variance component based tests are technically related to StructLMM. Variance component set tests, model  $\beta_{(G \times E)_s}$  in Eq. 2.101 as random effects following the distribution  $\beta_{(G \times E)_s} \sim \mathcal{N}(0, \tau^2)$  and score-based test statistics,  $Q$  and corresponding P values are derived similarly to the StructLMM interaction test (Section 2.2.5), setting  $\mathbf{K}_1 = \mathbf{T}\mathbf{T}^T$  and  $\mathbf{T} = [\mathbf{x}_1 \odot \mathbf{e}, \mathbf{x}_2 \odot \mathbf{e}, \dots, \mathbf{x}_S \odot \mathbf{e}]^T$ .

GESAT<sup>243</sup>, was the first variance component test proposed for interaction testing. To enable stable estimation of the marginal genetic effects for all  $S$  variants under

the null (instability may arise when large numbers of variants are included in the set, further compounded by the LD between them) ridge regression is used; by comparison in StructLMM where there are potentially large number of environments included in the set, a single variance component is used to estimate their marginal effect.

Whilst GESAT was designed for interaction set testing of common variants, a similar test by Chen *et al.*<sup>265</sup> was later developed to test for interactions with a set of rare variants. They proposed that the marginal genetic effects of rare variants,  $\beta_{G_s}$  in Eq. 2.101 are modelled as random effects following the distribution  $\beta_{G_s} \sim \mathcal{N}(0, \tau_1^2)$  such that a single variance component (with the same form as the tested genetic effect term in SKAT<sup>118</sup>) accounts for marginal genetic effects of the  $S$  variants under the null; this is analogous to the approach taken by StructLMM to account for marginal environment effects of  $L$  environments under the null. This proposal also enables a joint association test to be conducted, by setting  $\mathbf{K}_\rho = \rho \mathbf{G}\mathbf{G}^T + (1 - \rho)\mathbf{T}\mathbf{T}^T$ , where  $\mathbf{G} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_S]^T$  and using a similar testing procedure to SKAT-O<sup>123</sup> and the StructLMM joint association test described in this chapter.

An alternative to test for interaction effects with a set of rare variants is iSKAT<sup>244</sup>. iSKAT fits marginal genetic effects in a similar vein to GESAT; the key difference between the tests is that iSKAT uses MAF weighted ridge regression to fit the marginal genetic effects of the  $S$  variants, since these may not be comparable in magnitude across a set of rare variants in the same way that they are for common variants. A second key difference is that  $\beta_{(G \times E)}$  in Eq. 2.101 follows the distribution  $\beta_{(G \times E)} \sim \mathcal{N}(\mathbf{0}, \tau^2[\rho \mathbf{1}_S \mathbf{1}_S^T + (1 - \rho)\mathbf{I}_S])$ , where  $\beta_{(G \times E)} = [\beta_{(G \times E)_1}, \dots, \beta_{(G \times E)_S}]^T$ . This change, similar to the development of SKAT-O<sup>123</sup> from SKAT<sup>118</sup> for association tests, is to overcome the limitation that variance component tests may be underpowered when the majority of interactions in a set influence the phenotype with the same direction of effect; the iSKAT form of the random effect distribution for  $\beta_{(G \times E)}$  combines burden and variance component interaction tests to overcome this limitation. The score-based test statistics  $Q_\rho$ , the corresponding P values, the test statistic  $T$  and the corresponding P value are calculated similarly to the StructLMM joint association test, setting  $\mathbf{K}_\rho = \mathbf{T}\mathbf{R}_\rho\mathbf{T}^T$ , where  $\mathbf{R}_\rho = \rho \mathbf{1}_S \mathbf{1}_S^T + (1 - \rho)\mathbf{I}_S$ .

Whilst technically related to the StructLMM joint association test, the iSKAT<sup>244</sup> optimal test interpolates between fully correlated ( $\mathbf{T}\mathbf{1}_S \mathbf{1}_S^T \mathbf{T}^T$ ) and independent ( $\mathbf{T}\mathbf{I}_S \mathbf{T}^T$ ) variant effects after accounting for a single environmental exposure, where  $\rho$  can be interpreted as the average correlation of the variant effects across the  $S$

variants included in the set. The optimal test proposed by Chen *et al.*<sup>265</sup> interpolates between no G×E dependent ( $\mathbf{GI}_S\mathbf{G}^T$ ) and only G×E dependent ( $\mathbf{TI}_S\mathbf{T}^T$ ) variant effects, such that  $1-\rho$  can be interpreted as the average fraction of the variant effects explained by interactions with a single environment across the set of  $S$  considered variants. In comparison, the StructLMM optimal test interpolates between fully correlated ( $\text{diag}(\mathbf{x})\mathbf{1}_N\mathbf{1}_N^T\text{diag}(\mathbf{x})$ ) and environmentally dependent ( $\text{diag}(\mathbf{x})\mathbf{EI}_L\mathbf{E}^T\text{diag}(\mathbf{x})$ ) per-individual effects (in the special case that  $\mathbf{EE}^T = \mathbf{I}_N$ , there is no constraint on the similarity of the per-individual effects) and  $\rho$  can be interpreted as the fraction of the genetic effects explained by G×E at a single variant when accounting for multiple environmental exposures. Modelling of the latter required generalising the existing optimal testing procedure (as described in Section 2.2.5).

Genetic similarity regression is another set test based approach that has been proposed for interaction testing<sup>266–268</sup>, where trait similarity after accounting for covariates (including the environmental exposures themselves) is regressed on genetic similarity for a set of variants, taking into account environmental exposure. The model can be cast as:

$$\mathbb{E}(Z_{ij}) = aS_{ij} + bS_{ij} \times X_iX_j, i \neq j, \quad (2.102)$$

where  $Z_{ij}$  is the trait covariance of individuals  $i$  and  $j$ , obtained after accounting for covariates, i. e.  $Z_{ij} = (Y_i - \mu_i)(Y_j - \mu_j)$  where  $\mu_i$  is the subject specific mean.  $S_{ij}$  is the average genetic similarity over the  $S$  variants in the genetic set and  $X_i$  is the environmental exposure of individual  $i$ . A score test procedure, analogous to the StructLMM interaction and GESAT test, is used to test for interactions ( $b \neq 0$ ) and a score test under the assumption that  $b = d$  is used to test for joint associations ( $b = d \neq 0$ ). This joint association test is similar to that proposed by Chen *et al.*<sup>265</sup>, setting  $\rho = 0.5$  (i. e. no grid based search for optimal  $\rho$ ). Whilst not demonstrated, it is noted that the procedure can be extended to incorporate multiple environmental exposures by replacing  $X_i$  which describes a single environmental exposure, with  $\mathbf{X}_i$ , a  $1 \times L$  vector describing  $L$  environmental exposures and thus replacing  $X_iX_j$  by  $\mathbf{X}_i\mathbf{X}_j^T$ , but again under the assumption that marginal genetic and interaction effects explain the same amount of phenotypic variance, which is unlikely to be true. In addition, this method does not scale to large sample sizes.

Another interaction set based test that builds on existing multi-trait set tests (see Section 1.2.3)<sup>79</sup> is iSet<sup>80</sup>. By testing for associations between a set of genetic variants and multiple phenotypes, where the phenotype matrix comprises of phenotype measurements for the same trait in different environmental contexts, iSet can be

used to test for interaction effects with different categorical contexts. In addition, iSet can be used to distinguish whether the difference in SNP effect between the environments is consistent across all SNPs included within the set or whether the causal architecture differs between different contexts. This method is restricted to the analysis of categorical environments and is not scalable to large numbers of environmental variables or large sample sizes (it can handle at most tens of thousands of individuals).

A further variance component based interaction test is MAPIT<sup>269</sup>, designed to identify variants that are in epistasis with at least one other variant. Whilst conceptually related to StructLMM, the proposed testing procedure scales quadratically with the number of samples (StructLMM scales linearly) and thus is restricted to the analysis of cohorts with moderate sample size. In addition, this method is designed to test only for interactions, whilst StructLMM implements a joint association test.

All of the previously described interaction tests, were primarily designed to test for interactions with a single environmental exposure. The only interaction test, explicitly designed to test for interactions with multiple environments is that developed by Young *et al.*<sup>195</sup>, which focussed on a single variant (*rs1421085*) within the *FTO* locus. This multi-environment interaction test relies on a two step procedure. Briefly, in the first step they use a set of individuals to perform linear regression between BMI and environmental variables and use cross-validation to determine which environmental variables do not have any predictive power on BMI. Marginal variant and environment effects as well as  $G \times E$  interactions are then tested for at environments with predictive power, using a fixed effect multi-environment model similar to that described in Section 2.3.2, which as I will show in Section 2.4 is not as robust as StructLMM. In addition, there is no accompanying software available.

## 2.3 Comparison partners

In order to assess the performance of StructLMM as an interaction and joint association test, I compare results to a number of other approaches both in simulation experiments described in Section 2.4 and in applications to real data described in Chapter 3. These include single and multi environment models, as well as, linear models for the association test. A summary of these other methods is displayed in tabular form in Section 2.3.4.

### 2.3.1 Single environment models

As already described in Section 2.2.7, 1 df interaction tests between single environmental exposures and individual genetic variants are well established<sup>224</sup>. In addition, the same framework can be used for the 2 df joint association test<sup>218</sup>. There are three different single environment interaction models that I consider.

#### Single Environment model with Single environment additive effect

**(SingleEnv-Senv)** The standard single environment model is the same as that described by Eq. 2.99:

$$\mathbf{y} = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{(\mathbf{x} \odot \mathbf{e}_l)\beta_{G \times E_l}}_{\text{G} \times \text{E}} + \underbrace{\mathbf{e}_l\beta_{E_l}}_{\text{E}} + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (2.103)$$

with marginal form:

$$\mathbf{y} \sim \mathcal{N}\left(\underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{(\mathbf{x} \odot \mathbf{e}_l)\beta_{G \times E_l}}_{\text{G} \times \text{E}} + \underbrace{\mathbf{e}_l\beta_{E_l}}_{\text{E}}, \underbrace{\sigma_n^2 \mathbf{I}}_{\text{Noise}}\right), \quad (2.104)$$

where  $\mathbf{y}$  is an  $N \times 1$  phenotype vector for  $N$  individuals,  $\mathbf{W}$  is the  $N \times a$  fixed effect design matrix for  $a$  covariates,  $\boldsymbol{\alpha}$  is the  $a \times 1$  vector of their effect sizes,  $\mathbf{x}$  is an  $N \times 1$  genotype vector of the tested variant,  $\beta_G$  the corresponding genetic effect and  $\boldsymbol{\epsilon}$  the residuals (following the multivariate normal distribution  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}_N)$  as described in Eq. 2.8). This model can be used to assess the presence of  $G \times E$  with environment  $\mathbf{e}_l$  by testing  $\beta_{G \times E_l} \neq 0$  (1 df, SingleEnv-Senv-int) using a LRT. A joint P value that corresponds to the alternative hypothesis that at least one of  $L$  environments is participating in  $G \times E$  effects, can then be constructed by performing  $L$  tests followed by appropriate adjustment for multiple testing (I use Bonferroni). Similarly, a joint association test that accounts for  $G \times E$  effects due to single environments  $\mathbf{e}_l$  can be derived by testing  $[\beta_G, \beta_{G \times E_l}] \neq \mathbf{0}$  (2 df, SingleEnv-Senv), where again multiple environments and their corresponding tests can be combined using Bonferroni adjustment.

#### Single Environment model with multi-environment additive effect as

**Fixed effect (SingleEnv-Fenv)** When multiple environmental exposure measurements are available, the approach above can be extended by modelling

additive environment effects from multiple environments:

$$\mathbf{y} = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{(\mathbf{x} \odot \mathbf{e}_l)\beta_{G \times E_l}}_{\text{G} \times \text{E}} + \underbrace{\sum_{l=1}^L \mathbf{e}_l \beta_{E_l}}_{\text{E}} + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (2.105)$$

with marginal form:

$$\mathbf{y} \sim \mathcal{N}\left(\underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{(\mathbf{x} \odot \mathbf{e}_l)\beta_{G \times E_l}}_{\text{G} \times \text{E}} + \underbrace{\sum_{l=1}^L \mathbf{e}_l \beta_{E_l}}_{\text{E}}, \underbrace{\sigma_n^2 \mathbf{I}}_{\text{Noise}}\right). \quad (2.106)$$

This approach can improve the accuracy of the null model since additive effects of other potentially relevant environments are included. Again, the presence of interactions and associations can be assessed by testing  $\beta_{G \times E_l} \neq 0$  (SingleEnv-Fenv-int) and  $[\beta_G, \beta_{G \times E_l}] \neq \mathbf{0}$  (SingleEnv-Fenv) for each environment respectively, where again multiple environments can be combined using Bonferroni adjustment.

#### Single Environment model with multi-environment additive effect as

**Random effect (SingleEnv-Renv)** Alternatively, one can use a random effect to model multivariate additive environments, which is the approach taken in StructLMM. Since the null model of this approach is identical to that of StructLMM, this is the default single environment method used for comparison with StructLMM. Specifically, one can define an environmental covariance  $\boldsymbol{\Sigma}$  based on the observed environments as described in Section 2.2.4 and consider the model:

$$\mathbf{y} = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{(\mathbf{x} \odot \mathbf{e}_l)\beta_{G \times E_l}}_{\text{G} \times \text{E}} + \underbrace{\mathbf{u}}_{\text{E}} + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (2.107)$$

where  $\mathbf{u}$  follows the multivariate normal distribution  $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{\Sigma})$ , such that the marginal form is as follows:

$$\mathbf{y} \sim \mathcal{N}\left(\underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{(\mathbf{x} \odot \mathbf{e}_l)\beta_{G \times E_l}}_{\text{G} \times \text{E}}, \underbrace{\sigma_e^2 \boldsymbol{\Sigma}}_{\text{E}} + \underbrace{\sigma_n^2 \mathbf{I}}_{\text{Noise}}\right), \quad (2.108)$$

where again interaction (SingleEnv-Renv-int) and association tests (SingleEnv-Renv) can be implemented as described above.

### 2.3.2 Multi-environment models

An alternative strategy to model interaction effects between multiple environmental variables and a genotype is to use multiple fixed effects to account for the interactions. As already described in Section 2.2.4, StructLMM can be derived from this model by marginalising over the interaction effects. This approach requires an additional fixed effect for each additional environment that is tested for interactions such that the df of this test scale linearly with the number of environments. In comparison, for StructLMM the df are independent of the number of environments considered (1 df accounts for all tested environments), which can have calibration and power advantages as I will show in Section 2.4. Similarly, to StructLMM, the multi-environment models described here can be used to define both an interaction and a joint association test.

#### Multi-Environment model with multi-environment additive effect as

**Fixed effect (MultiEnv-Fenv)** Explicitly, denoting with  $\mathbf{e}_1, \dots, \mathbf{e}_L$  the  $N \times 1$  vectors for  $L$  environments, a fixed-effect based multi-environment model can be cast as:

$$\mathbf{y} = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{\sum_{l=1}^L (\mathbf{x} \odot \mathbf{e}_l)\beta_{G \times E_l}}_{\text{G} \times \text{E}} + \underbrace{\sum_{l=1}^L \mathbf{e}_l\beta_{E_l}}_{\text{E}} + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (2.109)$$

with marginal form:

$$\mathbf{y} \sim \mathcal{N}\left(\underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{\sum_{l=1}^L (\mathbf{x} \odot \mathbf{e}_l)\beta_{G \times E_l}}_{\text{G} \times \text{E}} + \underbrace{\sum_{l=1}^L \mathbf{e}_l\beta_{E_l}}_{\text{E}}, \underbrace{\sigma_n^2 \mathbf{I}}_{\text{Noise}}\right), \quad (2.110)$$

where both interactions due to  $L$  environmental variables and their additive effects are modelled as fixed effects. Within this model, the presence of interactions and associations can be assessed by testing  $[\beta_{G \times E_1}, \dots, \beta_{G \times E_L}] \neq \mathbf{0}$  ( $L$  df test) and  $[\beta_G, \beta_{G \times E_1}, \dots, \beta_{G \times E_L}] \neq \mathbf{0}$  ( $L + 1$  df test), respectively.

This test can be implemented using the LRT (named MultiEnv-Fenv-LRT-int and MultiEnv-Fenv-LRT for the interaction and association test, respectively).

Alternatively a score-based implementation can be employed (named MultiEnv-Fenv-Score-int and MultiEnv-Fenv-Score for the interaction and association test,

respectively), using Rao's score test statistic<sup>270</sup>.

$$\text{Rao's score test statistic} = \mathbf{U}_0^T \mathbf{I}_0^{-1} \mathbf{U}_0, \quad (2.111)$$

where  $\mathbf{U}_0$  is the gradient and  $\mathbf{I}_0$  the Fisher Information matrix with respect to the tested parameters, computed using MLE under the null<sup>†</sup>. Rao's score test statistic has an asymptotic chi-square distribution with the number of tested parameters as degrees of freedom.

### Multi-Environment model with multi-environment additive effect as

**Random effect (MultiEnv-Renv)** Alternatively, as for the single environment test, the multivariate additive environments can be modelled as a random effect instead of fixed effects, giving the following model:

$$\mathbf{y} = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{\sum_{l=1}^L (\mathbf{x} \odot \mathbf{e}_l) \beta_{G \times E_l}}_{\text{G} \times \text{E}} + \underbrace{\mathbf{u}}_{\text{E}} + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (2.115)$$

with marginal form:

$$\mathbf{y} \sim \mathcal{N} \left( \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{\sum_{l=1}^L (\mathbf{x} \odot \mathbf{e}_l) \beta_{G \times E_l}}_{\text{G} \times \text{E}}, \underbrace{\sigma_e^2 \boldsymbol{\Sigma}}_{\text{E}} + \underbrace{\sigma_n^2 \mathbf{I}}_{\text{noise}} \right). \quad (2.116)$$

Again as described above, both LR tests (MultiEnv-Renv-LRT-int, MultiEnv-Renv-LRT) and score tests (MultiEnv-Renv-Score-int, MultiEnv-Renv-Score) can be employed to test for interactions and associations respectively.

### 2.3.3 Linear association models

For the joint association test, standard linear models that assume constant genetic effect sizes in a population, can be used as an additional class of comparison methods.

---

<sup>†</sup>Specifically, for the test  $\boldsymbol{\beta} \neq \mathbf{0}$  in the Gaussian model

$$\mathbf{y} \sim \mathcal{N}(\mathbf{W}\boldsymbol{\alpha} + \mathbf{S}\boldsymbol{\beta}, \mathbf{H}), \quad (2.112)$$

$$\mathbf{U}_0 = (\mathbf{y} - \mathbf{W}\boldsymbol{\alpha}_0)^T \mathbf{H}_0^{-1} \mathbf{S}, \quad (2.113)$$

$$\mathbf{I}_0 = \mathbf{S}^T \mathbf{H}_0^{-1} \mathbf{S}, \quad (2.114)$$

where  $\boldsymbol{\alpha}_0$  and  $\mathbf{H}_0$  are MLE of  $\boldsymbol{\alpha}$  and  $\mathbf{H}$  under the null model.



**Linear model (LM)** The standard linear model is cast as:

$$\mathbf{y} = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (2.117)$$

with marginal form:

$$\mathbf{y} \sim \mathcal{N}(\underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}}, \underbrace{\sigma_n^2 \mathbf{I}}_{\text{Noise}}). \quad (2.118)$$

A 1 df LR test  $\beta_G \neq 0$  is used to test for association.

**Linear model with multivariate additive environment effects modelled as Fixed effect (LM-Fenv)** When  $L$  environmental exposure measurements are available, multivariate additive environment effects can be accounted for using fixed effects as follows:

$$\mathbf{y} = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{\sum_{l=1}^L \mathbf{e}_l \beta_{E_l}}_{\text{E}} + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (2.119)$$

with marginal form:

$$\mathbf{y} \sim \mathcal{N}(\underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{\sum_{l=1}^L \mathbf{e}_l \beta_{E_l}}_{\text{E}}, \underbrace{\sigma_n^2 \mathbf{I}}_{\text{Noise}}). \quad (2.120)$$

**Linear model with multivariate additive environment effects modelled as Random effect (LMM-Renv)** Alternatively, similarly to the single and multi-environment models, the multivariate additive environmental effects can be modelled as random:

$$\mathbf{y} = \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{\mathbf{u}}_{\text{E}} + \underbrace{\boldsymbol{\epsilon}}_{\text{Noise}}, \quad (2.121)$$

with marginal form:

$$\mathbf{y} \sim \mathcal{N}(\underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{Covariates}} + \underbrace{\mathbf{x}\beta_G}_{\text{G}}, \underbrace{\sigma_e^2 \boldsymbol{\Sigma}}_{\text{E}} + \underbrace{\sigma_n^2 \mathbf{I}}_{\text{Noise}}). \quad (2.122)$$

Since this model is identical to StructLMM under the null, this is the default comparison partner used.

### 2.3.4 Summary of comparison partners

Table 2.2, summarises the modelling choice for the interaction and additive environment terms, as well as the number of degrees of freedom (based on testing  $L$  environments) for the various comparison methods described above.

Method name	GxE	Additive E	Test type	Number parameters for additive E	Number df Interaction test	Number df association test
Multi-environment tests						
<b>StructLMM</b>	<b>Random</b>	<b>Random</b>	<b>Score</b>	<b>1</b>	<b>1</b>	<b>2</b>
MultiEnv-Renv-LRT	Fixed	Random	LRT	1	$L$	$L + 1$
MultiEnv-Fenv-LRT	Fixed	Fixed	LRT	$L$	$L$	$L + 1$
MultiEnv-Renv-Score	Fixed	Random	Score	1	$L$	$L + 1$
MultiEnv-Fenv-Score	Fixed	Fixed	Score	$L$	$L$	$L + 1$
Single-environment tests						
<b>SingleEnv-Renv</b>	<b>Fixed</b>	<b>Random</b>	<b>LRT</b>	<b>1</b>	<b>1</b>	<b>2</b>
SingleEnv-Fenv	Fixed	Fixed	LRT	$L$	1	2
SingleEnv-Senv	Fixed	Fixed	LRT	1	1	2
Linear association tests						
<b>LMM-Renv</b>	<b>None</b>	<b>Random</b>	<b>LRT</b>	<b>1</b>	<b>NA</b>	<b>1</b>
LM-Fenv	None	Fixed	LRT	$L$	NA	1
LM	None	None	LRT	0	NA	1

**Table 2.2 Comparison methods considered** | Summary of the comparison methods used throughout this thesis, showing the method name (column 1) used for association testing and for interaction testing ‘-int’ is appended to the listed names, whether a random or fixed effect term is used to model  $G \times E$  under the alternative hypothesis (column 2), whether a random or fixed effect term is used to model the additive environment (column 3), the statistical test used to assess the alternative hypothesis (column 4), how many parameters are used to model the additive environment (column 5), the number of additional parameters used to model the alternative hypothesis versus the null hypothesis for the interaction and association test (columns 6 and 7 respectively). The number of model parameters in the final three columns are based on testing  $L$  environments for interaction effects. The tests are grouped into multi-environment tests, single environment tests, and linear association tests. Methods that are used as the default comparison partners are in bold.

## 2.4 Simulation experiments

Simulation experiments were used to show that StructLMM is calibrated, and then to demonstrate the power advantages of StructLMM compared to other methods for the interaction and association test (Section 2.3). In this section, I will first describe the data used for these experiments, the simulation procedure used to generate phenotypes and the methods used to assess calibration and statistical power. I will then show calibration and power results for some general settings, followed by results for simulation experiments that explicitly examine the effect of certain environmental properties.

### 2.4.1 Simulation data

#### Genotype data

For the simulation procedure, genotypes were derived based on the 372 European individuals (from CEU, FIN, GBR, IBS and TSI ancestries) from the 1000 Genomes Project phase 1<sup>94</sup> (1,092 individuals in total). The generation of a set of  $N$  samples of European ancestry follows the procedure proposed by Loh *et al.*<sup>77</sup>, (see also<sup>79,80</sup>). Briefly, 10 samples from the original data were selected (‘ancestors’) and then blocks of 1,000 SNPs ‘inherited’ to create new individuals; this procedure retains LD structure and using 10 ancestors, results in realistic population structure without the inclusion of close relatives. Populations of size  $N = \{1,000, 2,000, 5,000\}$  were generated, with  $N = 5,000$ , the default setting used for simulation experiments (see Table 2.3). Only variants with a MAF > 2%, were used in the simulation experiments and the variants were mean centred and standardised.

#### Environmental exposure data

The environmental covariates in this chapter were based on the UK Biobank interim data release (Application 14069)<sup>215,271</sup> to mimic realistic environmental distributions. The interim release contained genotype data for 152,729 samples of the total ~500,000 samples available in the full release (see Section 1.2.2 for further description of the UK Biobank cohort and further details of this interim release are available at [http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank\\_genotyping\\_QC\\_documentation-web.pdf](http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UKBiobank_genotyping_QC_documentation-web.pdf)). Thirty-two environmental variables, 9 ordinal dietary variables (‘Oily fish intake’, ‘Non-oily fish intake’, ‘Processed meat intake’,

‘Poultry intake’, ‘Beef intake’, ‘Lamb/mutton intake’, ‘Pork intake’, ‘Cheese intake’ and ‘Salt added to food’), three continuous dietary variables (‘Cooked vegetable intake’, ‘Bread intake’, ‘Tea intake’), three physical activity variables (‘Number of days/week walked 10+ minutes’, ‘Number of days/week of moderate physical activity 10+ minutes’, ‘Number of days/week of vigorous physical activity 10+ minutes’), ‘Alcohol intake frequency’, ‘Sleep duration’, ‘Sleep duration residuals squared’, ‘Townsend deprivation index’, ‘Smoking status’, ‘Time spent watching television’, ‘Usual walking pace’, ‘Frequency of friend/family visits’, ‘Time spend outdoors in summer’, ‘Time spent outdoors in winter’, ‘Time spent using computer’, ‘Nap during day’, ‘Overall health rating’ and five pollutant measures (‘Nitrogen dioxide air pollution 2010’, ‘Nitrogen oxides air pollution 2010’, ‘Traffic intensity on the nearest major road’, ‘Average daytime sound level of noise pollution’, ‘Average evening sound level of noise pollution’), using data from ‘Instance 0’ were selected (see UK Biobank Data Showcase, <http://biobank.ctsu.ox.ac.uk/crystal/> for further information). I first removed the 480 individuals that were flagged as poor quality samples (UK Biobank field ‘22010’). Then following Young *et al.*<sup>195</sup>, for the three continuous dietary and five pollutant variables, I removed values exceeding the 99th percentile. For ‘Sleep duration’, I removed the top and bottom percentiles and for each individual, calculated the squared deviations from the mean sleep duration, creating environmental variable, ‘Squared sleep duration res.’ (33<sup>rd</sup> environmental variable). For the four variables (‘Time spent watching television’, ‘Time spent using computer’, ‘Time spend outdoors in summer’, ‘Time spent outdoors in winter’), less than 0.5 hours of was encoded as 0.5 and I excluded individuals in the upper and lower percentile. Smoking status was converted from three categories, ‘Current’, ‘Previous’ and ‘Never’ to a binary variable, ‘Ever’ and ‘Never’. I then removed individuals with any missing environmental data, leaving a total of 92,600 individuals. Finally, I removed any remaining individuals that had 3<sup>rd</sup> degree or closer relative based on UK Biobank field ‘22011’, using the field entitled ‘Recommended genomic analysis exclusions 0.0’ and kept only individuals defined as ‘White British’ using UK Biobank field ‘22006’ (individuals who identified as ‘White British’ and have a similar genetic ancestry based on principal component analysis of the genotype data), resulting in 70,282 environmental profiles available for the simulation procedures.

These 33 environmental variables were age and gender adjusted, to allow for interaction effects that are gender specific or age dependent, as described in Section 2.2.4, and age itself added to the environmental matrix,  $\mathbf{E}$ , resulting in a total of 100 environmental covariates.  $N$  of the 70,282 profiles were randomly selected and assigned to the generated genotypes (described in Section 2.4.1) and  $L$  of the 100

environments selected at random (the default setting is  $L = 60$ , see Table 2.3).

As described in Section 2.4.5, for the simulation experiments conducted to examine the effect of specific environmental properties, these UK Biobank environments were used to generate environmental variables with the desired characteristics.

Each environmental variable (or the subsequently generated environmental variables as described in Section 2.4.5) was mean centred and standardised, followed by a rescaling by factor  $\frac{1}{\sqrt{L}}$  (a standard linear covariance procedure as described in Section 2.2.4).

## 2.4.2 Phenotype simulation procedure

The  $N \times 1$  phenotype vector,  $\mathbf{y}$ , was simulated as the sum of a persistent genetic contribution ( $\mathbf{g}$ ), a G×E contribution ( $\mathbf{i}$ ), an additive environmental contribution ( $\mathbf{e}$ ), a population structure contribution ( $\mathbf{u}$ ) and a random noise contribution ( $\boldsymbol{\epsilon}$ ), such that:

$$\mathbf{y} = \mathbf{g} + \mathbf{i} + \mathbf{e} + \mathbf{u} + \boldsymbol{\epsilon}. \quad (2.123)$$

Each of these terms is simulated as follows:

- **Persistent genetic effect,  $\mathbf{g}$ :** A genetic variant is selected at random and then rescaled by a factor of  $\frac{v_g(1-\rho)}{\text{std}(\text{variant})}$ , such that the resulting vector has sample variance  $v_g(1-\rho)$ .  $\text{std}(\text{variant})$  is the standard deviation of the selected variant,  $v_g$  is the fraction of the phenotypic variance explained by genetics (i.e. the combination of persistent and G×E effects) and  $\rho$ , the fraction of the total genetic contribution explained by G×E effects. By default,  $v_g = 0.006$  and  $\rho = 0.7$  (see Table 2.3 for other values of these parameters used).
- **G×E effect,  $\mathbf{i}$ :** First a fraction ( $\pi$ ) of the  $L$  environmental variables were selected (where  $L$  is as described in Section 2.4.1). Then the same genetic variant that was used to simulate the persistent genetic effect was multiplied element wise by  $\{1, -1\}$  selected at random and further multiplied with the selected environmental variables. That is  $\mathbf{i} = \mathbf{E}_{\text{subset}} \boldsymbol{\beta}_{\text{G} \times \text{E}} \odot \mathbf{x}$ , where  $\mathbf{E}_{\text{subset}}$  is an environmental matrix comprised of the selected environmental variables (i.e. it is a subset of the columns of  $N \times L$  environmental matrix,  $\mathbf{E}$  described in Section 2.4.1).  $\boldsymbol{\beta}_{\text{G} \times \text{E}}$  is a vector of the G×E effects for each of the selected environmental variables, randomly selected from  $\{1, -1\}$ , such that selected environments contribute equally to the generated interaction term.  $\mathbf{i}$  is then renormalised such that the sample variance is  $v_g \rho$ . By default  $\pi = 0.5$  of the

$L$  environments are selected such that by default 30 environments are used to simulate the  $G \times E$  interaction effect (see Table 2.3 for other values of the parameters used).

In addition to considering settings where the environments contributing to the  $G \times E$  effect are a random subset of those randomly selected to generate the additive environment effect (fraction  $\pi$  of the  $L$  environments; see next bullet point), settings in which  $L_{\text{unobs}}$  additional environments (from the 100 total;  $L_{\text{unobs}}$  defines the number of additional environments) to the  $L$  already selected were considered to generate an environmental matrix  $\mathbf{E}_{\text{superset}}$ , which contains columns with additional environmental variables compared to environmental matrix  $\mathbf{E}$  described in Section 2.4.1. This corresponds to a setting where an environmental driver that is not observed or measured but is somewhat correlated with the observed environmental variables, contributes to the interaction effect, a scenario that is likely to occur in reality.

For all experiments (both when a fraction  $\pi$  of the  $L$  environments are used and when  $L_{\text{unobs}}$  environments in addition to the  $L$  environments are used to simulate the  $G \times E$  effect), the original  $L$  environments are used to simulate the additive environment effect (see next bullet point). In all cases the testing models use the  $L$  environments for both the  $G \times E$  and the additive environment term (see Section 2.4.3). This strategy results in no model mismatch of the additive environment term.

- **Additive environmental effect,  $\mathbf{e}$ :** For each of the  $L$  environmental variables (selected as described in Section 2.4.1), an effect  $\beta_{E_l}$  is randomly generated from  $\beta_{E_l} \sim \mathcal{N}(0, 1)$  and then  $\mathbf{e}$  is calculated as  $\mathbf{e} = \sum_{l=1}^L \mathbf{e}_l \beta_{E_l}$ .  $\mathbf{e}$  is then renormalised to have sample variance  $v_e$ , for which the default setting is  $v_e = 0.2$  (see Table 2.3 for other values of this parameter used).
- **Population structure,  $\mathbf{u}$ :** The first ten principal components from the kinship matrix (see Section 1.1.6 for a description) of the genotype data are used to generate population structure. The effect,  $\beta_{\text{pop}_p}$ , of each principal component ( $\text{PC}_p$ ), was generated randomly as  $\beta_{\text{pop}_p} \sim \mathcal{N}(0, 1)$  and then  $\mathbf{u}$  is calculated as  $\mathbf{u} = \sum_{p=1}^{10} \text{PC}_p \beta_{\text{pop}_p}$ .  $\mathbf{u}$  is then renormalised to have sample variance  $v_{\text{pop}}$ , which is always set to  $v_{\text{pop}} = 0.4$ .
- **Random noise,  $\boldsymbol{\epsilon}$ :** An  $N \times 1$  noise vector is generated as  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_N)$  and then rescaled to have sample variance  $1 - v_g - v_e - v_{\text{pop}}$ .

General simulation parameters										
$N$	-	-	-	1,000	2,000	<b>5,000</b>	-	-	-	-
$L$	-	2	10	20	40	<b>60</b>	80	100	-	-
$\pi$	-	-	0.033	0.167	0.333	<b>0.50</b>	0.667	1	-	-
$v_g$	-	-	-	-	-	<b>0.006</b>	-	-	-	-
$\rho$	-	0.0	0.1	0.3	0.5	<b>0.7</b>	0.8	0.9	1.0	-
$v_e$	-	-	-	0	0.1	<b>0.2</b>	0.3	0.4	-	-
$v_{\text{pop}}$	-	-	-	-	-	<b>0.4</b>	-	-	-	-
$L_{\text{unobs}}$	-	-	-	-	-	<b>0</b>	10	20	30	40
Skewed (Gamma-distributed) environments										
$k$	-	100	5	3	2	1	0.5	0.2	0.1	-
Binary environments										
$f_B$	-	0	0.25	0.50	0.75	<b>1</b>	-	-	-	-
$\nu$	-	0.50	0.20	0.10	0.05	<b>0.02</b>	0.01	0.005	-	-
Heritable environments										
$r^2$	-	0	0.25	0.50	0.75	<b>1</b>	-	-	-	-
$v_x$	-	0.50	0.20	0.10	0.05	<b>0.02</b>	0.01	0.005	-	-

**Table 2.3 Parameters used for simulation experiments** | Shown are the parameter values considered in calibration and power experiments (see Sections 2.4.4 and 2.4.5).  $N$  is the population sample size,  $L$  the number of selected environmental variables used to generate the additive environment effect,  $\pi$  the fraction of the  $L$  environmental variables selected for the generation of the interaction effect,  $v_g$  the fraction of the phenotypic variance explained by the total genetic effect (G+G×E),  $\rho$  the fraction of the total genetic effect explained by G×E effects,  $v_e$  the fraction of the phenotypic variance explained by additive environment effects,  $v_{\text{pop}}$  the fraction of the phenotypic variance explained by population structure and  $L_{\text{unobs}}$  the number of environments selected in addition to the  $L$  already selected. In simulation settings that examine the effect of certain environmental properties (described in Section 2.4.5), additional parameters are also defined. For skewed environments, the shape parameter  $k$  of the Gamma distribution is varied; for binary environments, both the fraction of environments  $f_B$  that are binary and the event frequency  $\nu$  are varied; and for heritable environments, both the LD,  $r^2$ , between the variant that is associated with the environment and that used to generate the persistent and G×E effects and the average fraction of the environments explained by the variant  $v_x$  (heritability) are varied. Unless otherwise specified, the parameters displayed in bold are those used by default.

## 2.4.3 Simulation approach

### Simulation model settings

Once the phenotypes were simulated, various models (see Section 2.3 for details on the models compared) were used to test for interactions and associations for windows

of variants (including the variant driving the genetic effect and that associated with the environments in the heritable setting described in Section 2.4.5). For all models, the included covariates were a mean intercept term (vector of ones) and the ten principal components used to generate  $\mathbf{u}$  (as described in Section 2.4.2). The additive environment and  $G \times E$  terms use the  $L$  selected environments (as described in Section 2.4.2 i.e. mean centred and standardised and then rescaled by a factor  $\frac{1}{\sqrt{L}}$ , such that  $\Sigma$  used in testing is a standard linear covariance matrix as described in Section 2.2.4).

## Calibration method

Calibration was assessed by splitting the 103,527 variants on chromosome 21 into 100 chunks. Phenotypes were generated (as described in Section 2.4.2) for each chunk, either with no genetic effects ( $v_g = 0$ ) or with persistent genetic effects driven by one randomly selected causal variant ( $v_g = 0.006$ ,  $\rho = 0$ ). This was repeated 100 times such that there were a total of  $\sim 10^7$  P values.

All generated P values were pooled and QQ plots of the expected negative log P values (x-axis) versus the observed negative log P values (y-axis) were generated and inflation parameters,  $\lambda_{GC} = \frac{\log_{10}(m)}{\log_{10}(0.5)}$ , where  $m$  is the median P value over all variants tested, were calculated. If  $\lambda_{GC} \approx 1$ , then the method is deemed calibrated, with  $\lambda_{GC}$  much greater than 1 indicative of inflation and  $\lambda_{GC}$  much smaller than 1 indicative of deflation.

## Statistical power method

Statistical power was assessed by selecting genomic chunks from chromosome 21 of  $\sim 2$  Mb (3,000 SNPs). Phenotypes were simulated (as described in Section 2.4.2) with each chunk containing one causal variant. This was repeated 1,000 times for each simulation setting.

For each experiment, a score of 1 (successful identification of the causal variant) was assigned if the simulated causal variant or a variant in LD,  $r^2 \geq 0.8$  had P value  $< 0.01$  after Bonferroni correction for the number of tests, corresponding to a 1% FWER. If this was not the case, 0 (unsuccessful identification of the causal variant) was assigned. Power is then defined as the average score over the 1,000 repeat experiments, resulting in a value lying between 0 and 1, where 0 corresponds to the case that the causal variant is not identified in any experiment and 1, the causal



variant is identified in all experiments.

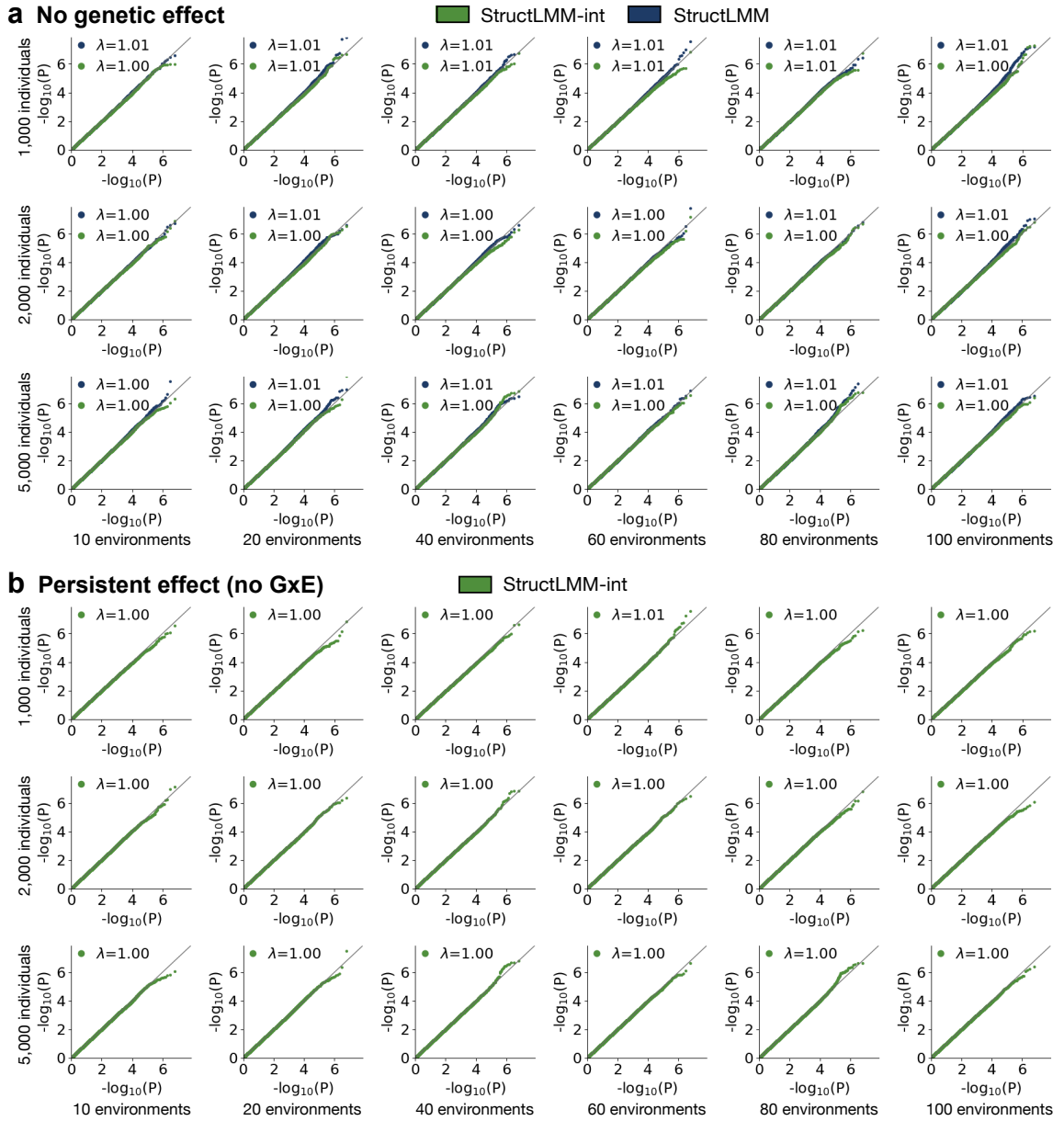
An alternative rank based method for assessing model performance, was used to compare methods that were not always calibrated. P values per experiment were ordered from most to least significant (smallest to largest) and the cumulative true positive rate (TPR) and false positive rate (FPR) calculated, where again true positives were defined as the simulated causal or variants in LD,  $r^2 \geq 0.8$  and false positive otherwise. These cumulative rates were averaged over the repeat experiments and then the area under the curve (AUC, where the average cumulative FPR is plotted on the x-axis and average cumulative TPR plotted on the y-axis; see Fawcett *et al.*<sup>272</sup> for further details) calculated using a FPR limit of 10%. This value was then normalised, such that 0 corresponds to chance performance and 1 to a perfect performance (here all variants in LD,  $r^2 \geq 0.8$  would have smaller P values than all variants not in LD,  $r^2 < 0.8$ ).

## 2.4.4 Simulation results

In this section, I outline the calibration and power results for some general simulation settings.

### StructLMM calibration results

The statistical calibration of the StructLMM interaction test (StructLMM-int) and StructLMM joint association test were assessed when no causal variants were simulated ( $v_g = 0$ , Fig. 2.3a) and StructLMM-int when persistent genetic effects were simulated ( $v_g = 0.006$ ,  $\rho = 0$ , Fig. 2.3b), with increasing numbers of environmental variables contributing to the additive environmental effect (there was no simulated G×E effect;  $\pi = 100\%$ ) for sample sizes of 1,000, 2,000 and 5,000 individuals. In all cases, StructLMM interaction and association tests were calibrated.



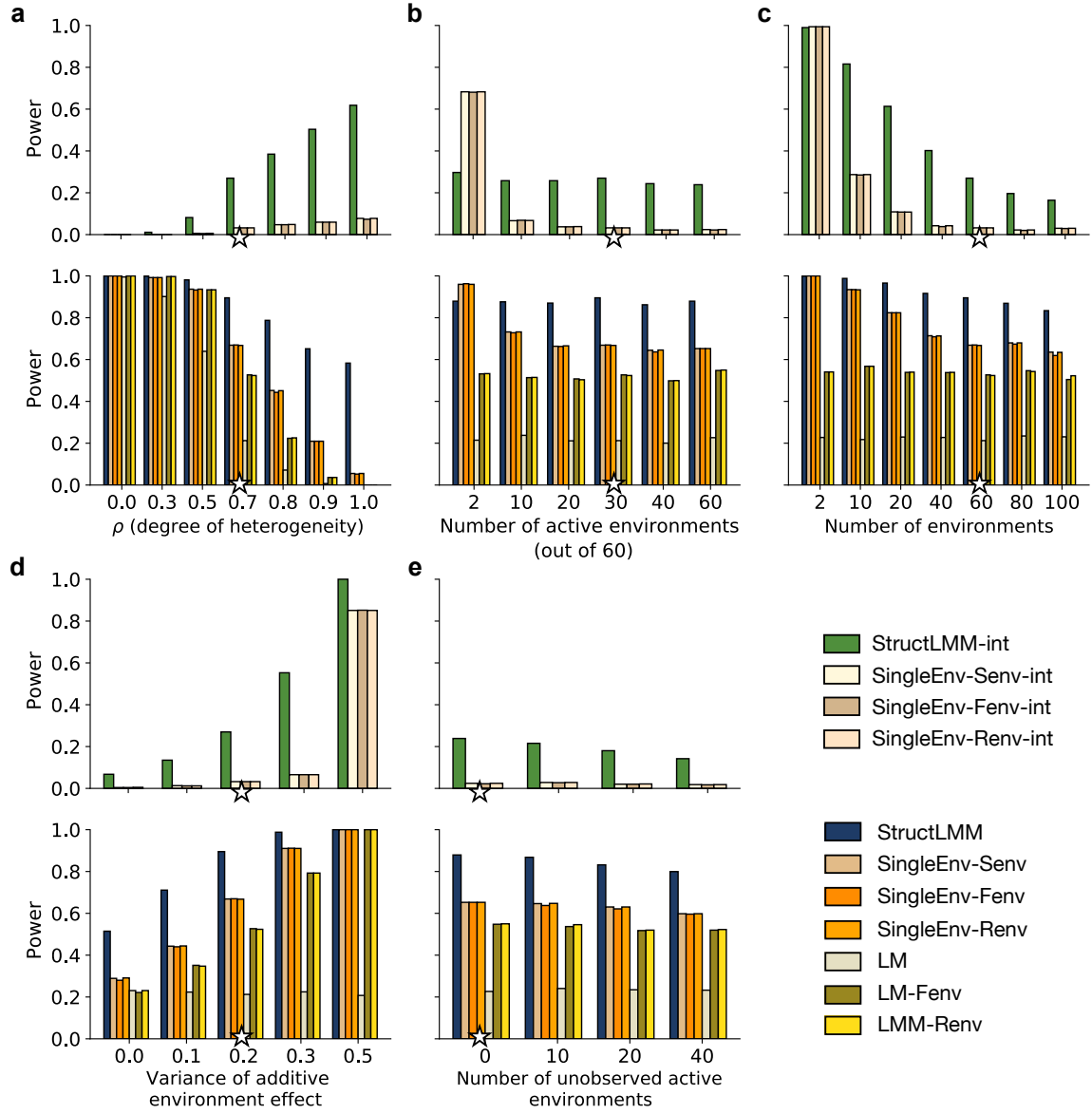
**Fig. 2.3 Calibration of StructLMM interaction and association test** | (a) QQ plots of negative log P values from the StructLMM interaction test (StructLMM-int, green) and StructLMM association test (blue) either simulating no genetic effects (no G, no G×E,  $v_g = 0$ ) or (b) simulating persistent genetic effects (no G×E,  $v_g = 0.006$ ,  $\rho = 0$ ; only StructLMM-int). The genomic inflation factors  $\lambda_{GC}$  (denoted by  $\lambda$ ) for the interaction and association tests are displayed in the top left of each plot. From top to bottom: increasing sample sizes ( $N$ ) of a synthetic population based on the European population from the 1000 Genomes project: 1,000 individuals, 2,000 individuals and 5,000 individuals. From left to right: increasing numbers of environmental variables ( $L$ ) used for simulations and tests ( $\pi = 100\%$ ): 10 to 100.

## Comparison to other methods

The statistical power of StructLMM interaction and association tests were compared to other interaction and association methods (see Section 2.3 for details on the models compared). Initially, for interaction tests, StructLMM-int and all implementations of the single-environment 1 df interaction tests (Table 2.2) were considered and for association tests, StructLMM, all 2 df single environment association tests (Table 2.2) and all implementations of the linear models (Table 2.2) were compared.

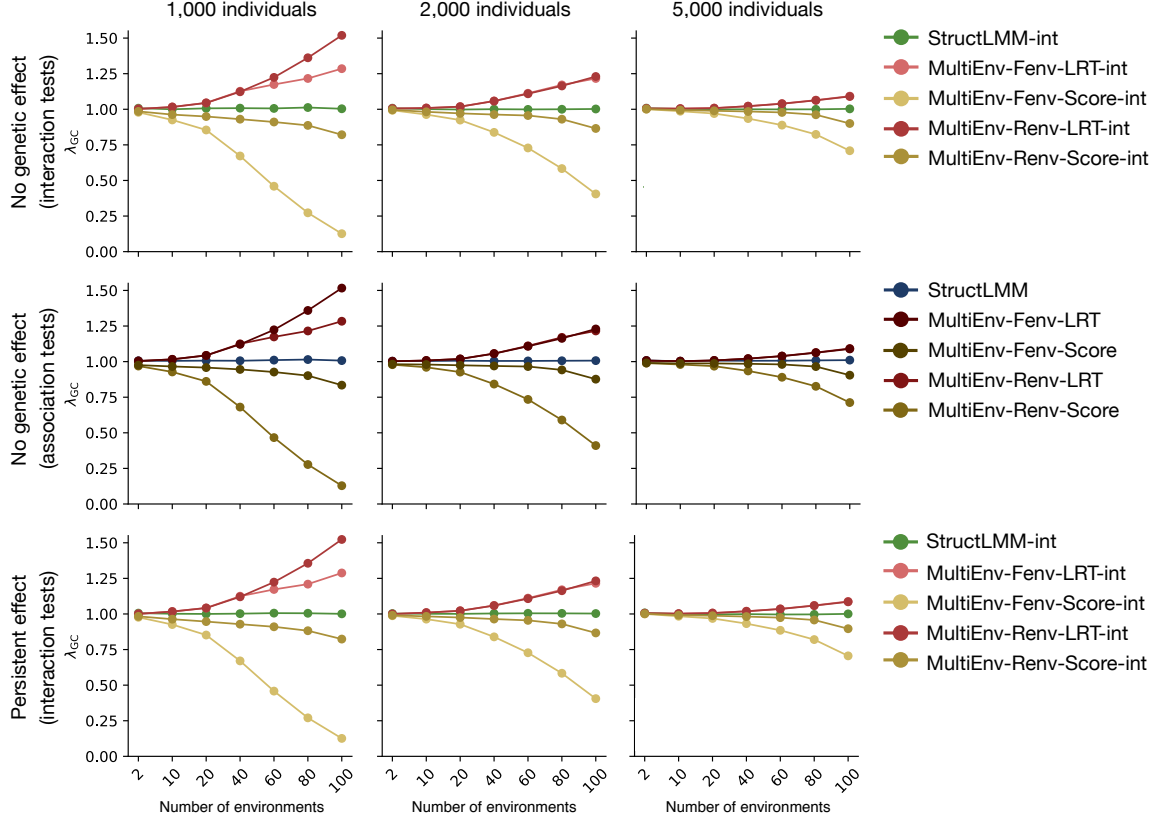
First, the proportion of the genetic effect driven by  $G \times E$  ( $\rho$ ) was increased from 0 to 1 (Fig. 2.4a). This resulted in increased power of the interaction tests and decreased power of the association tests as expected, noting that the StructLMM models increasingly outperformed other considered methods. Other default parameters were then varied, including the percentage ( $\pi$ ) of the 60 tested environments that contribute to the  $G \times E$  effect (Fig. 2.4b), the number of environments ( $L$ , using  $\pi = 50\%$  to simulate  $G \times E$  effects and  $\pi = 100\%$  to simulate additive environment effects) using  $L$  environments when testing (Fig. 2.4c), and the amount of phenotypic variance explained by additive environment effects ( $v_e$ , Fig. 2.4d). Finally, the scenario in which environments, in addition to the 60 observed (and tested), were used to simulate the  $G \times E$  effect ( $L_{\text{unobs}}$ , Fig. 2.4e), corresponding to the case where the true  $G \times E$  environmental drivers are unmeasured and therefore cannot be included when testing for interaction and association effects. In all settings StructLMM performed better than the other baseline methods. Additionally, it can be seen that the modelling choice for the additive environment term made little difference to the results; thus in the remaining simulation settings and in applications to real data (Chapter 3), a random additive environment effect term is used (see methods highlighted in bold in Table 2.2), such that the null model is identical to that of StructLMM tests.

As already mentioned in Section 2.2.4 and 2.3.2, StructLMM can be derived by marginalising over the multiple  $G \times E$  terms present in a fixed effect framework. Therefore, as a further comparison, StructLMM was compared to other implementations of multi-environment fixed effect methods (Table 2.2). These models were not always calibrated, in particular when the number of environments,  $L$ , relative to the number of samples,  $N$ , was large (probably due to the high number of df), finding that the LRT was inflated whilst the score test was deflated (Fig. 2.5). Therefore, to compare the performance of these models, a rank based AUC method was used (see Section 2.4.3 for details), when varying the fraction of genetic effects



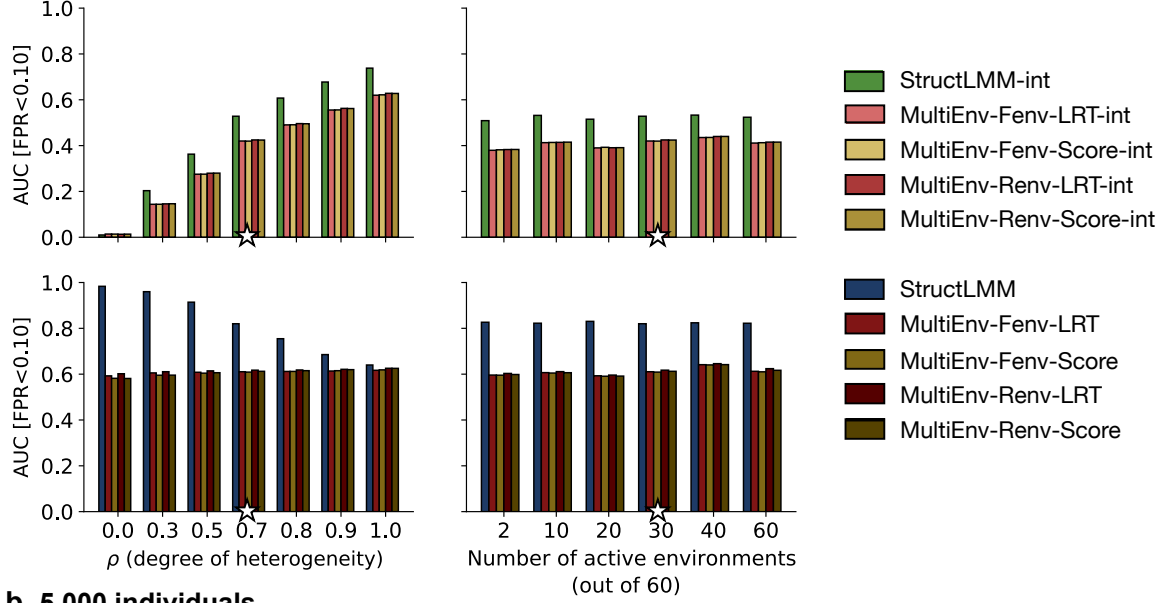
**Fig. 2.4 Power comparison of different methods** | Power comparison of alternative methods for detecting interactions (top panels) and associations (bottom panels) based on simulated data, increasing (a) the fraction of the genetic variance explained by  $G \times E$  ( $\rho$ ), (b) the number of environments with non-zero  $G \times E$  effects ( $\pi$ ), (c) the total number of environments ( $L$ ,  $\pi = 50\%$  contributing to  $G \times E$  effects), (d) the fraction of variance explained by additive environment effects ( $v_e$ ) and (e) the number of environments that contribute to  $G \times E$  but are not used (observed) for the respective tests ( $L_{\text{unobs}}$ ). In the top panels, considered are the StructLMM interaction test (StructLMM-int) and alternative implementations of the single environment interaction test (Table 2.2). In the bottom panels, considered are the StructLMM association test, alternative implementations of the 2 df fixed effect tests that test for associations whilst accounting for possible heterogeneity in the variant effect due to interactions with a single environment (Table 2.2), and all implementations of linear models that test for persistent effects (Table 2.2). Models are assessed in terms of power (FWER < 1%) for detecting simulated causal variants. Stars denote default values of genetic parameters, which are retained when varying other parameters (see Table 2.3).

driven by  $G \times E$  ( $\rho$ ) and the percentage ( $\pi$ ) of the 60 tested environments that contribute to the  $G \times E$  effect, for two sample sizes ( $N = 2,000, 5,000$ ). As well as retaining calibration, StructLMM is also better powered than these other multivariate implementations (Fig. 2.6).

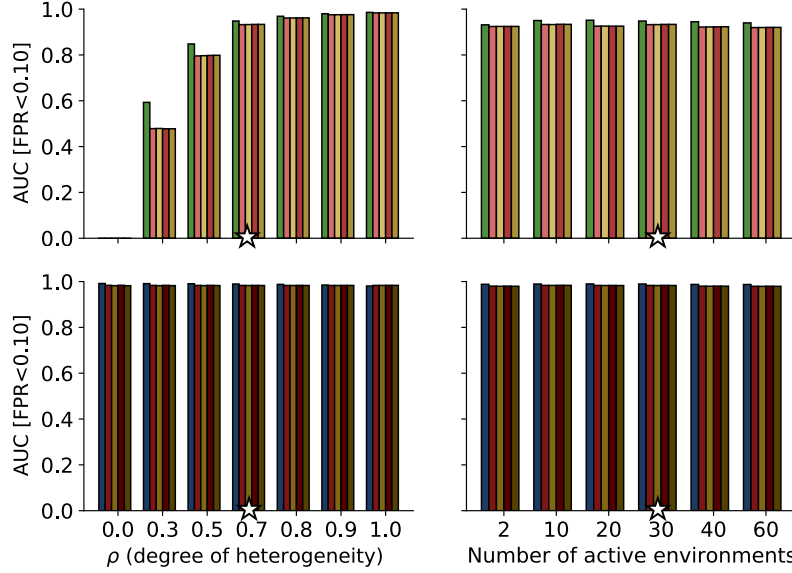


**Fig. 2.5 Calibration comparison of different implementations of multi-environment tests** | Genomic inflation factor  $\lambda_{GC}$  of P values from StructLMM and alternative implementations of multi-environment tests based on fixed effects (Table 2.2), for different numbers of environmental variables ( $L$ ,  $\pi = 100\%$ ; x-axis) and for increasing sample size ( $N$ , left to right). Shown are results from StructLMM-int and multi-environment fixed effect interaction tests (row 1), and equivalent association tests (row 2) when no genetic effects are simulated (no  $G$ , no  $G \times E$ ,  $v_g = 0$ ). Row 3 depicts results from StructLMM-int and multi-environment fixed effect interaction tests when persistent genetic effects are simulated (no  $G \times E$ ,  $v_g = 0.006$ ,  $\rho = 0$ ). Multi-environment fixed effect models using LR tests yielded inflated test statistics (inflation factors  $\lambda_{GC} > 1$ ) for large numbers of environmental factors in relation to the sample size, whilst score tests yielded deflated test statistics (inflation factors  $\lambda_{GC} < 1$ ) for the corresponding settings, whereas StructLMM was calibrated in all settings.

### a 2,000 individuals



### b 5,000 individuals



**Fig. 2.6 Power comparison of different implementations of multi-environment tests** | Assessment of performance of alternative methods for detecting interactions (top panels) and associations (bottom panels) based on simulated data, for the same simulation settings as considered in Fig. 2.4a, b for two sample sizes:  $N = 2,000$  (**a**) and  $N = 5,000$  (**b**). Compared are StructLMM-int and all other implementations of multi-environment interaction tests (Table 2.2, top panels) and StructLMM and all other implementations of multi-environment association tests (Table 2.2, bottom panels). As the fixed effect tests are not always calibrated (see Fig. 2.5), shown are model performance values as assessed by the area under the curve (AUC, in the range  $0 < \text{FPR} < 0.10$ , normalised such that 0 corresponds to chance performance and 1 to an ideal model).

## 2.4.5 Simulation results examining the effect of specific environmental properties

In this section, I outline the calibration and power results for simulation settings, in which the environmental variables are generated to have specific properties. I will first describe the general method used to simulate these synthetic environments and then explore settings when the environments are skewed and/or binary and settings when the environmental variables are themselves heritable.

### Generating synthetic environments

For the remaining simulation settings, a matrix normal distribution (see Gupta *et al.*<sup>273</sup> for details) was used to generate the environment matrix, based on the sample and environment covariances for the selected UK Biobank environments (as described in Section 2.4.1). Explicitly:

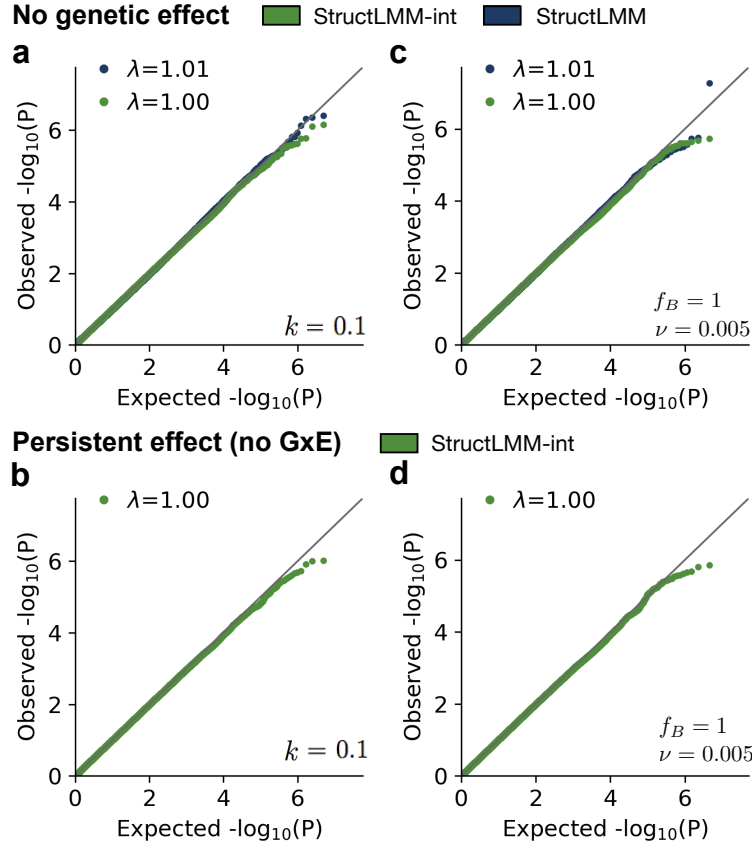
$$\mathbf{E}_L \sim \mathcal{MN}_{N \times L}(\mathbf{0}, \mathbf{R}, \mathbf{C}) \quad (2.124)$$

where the  $N \times L$  environment matrix  $\mathbf{E}_L$  is generated from the matrix normal distribution, with  $\mathbf{R}$  the  $N \times N$  sample covariance matrix describing the covariance between pairs of individuals across the  $L$  environments and  $\mathbf{C}$  the  $L \times L$  environmental covariance matrix describing the covariance between pairs of environments across the  $N$  individuals.

### Skewed and binary environments

For the skewed and binary environment simulation experiments  $\mathbf{C}$  and  $\mathbf{R}$  were calculated directly from the randomly selected UK Biobank environment data to generate new continuous environments with realistic correlation structure. To assess the impact of using non-gaussian environments, all environments were rank-inverse transformed to a Gamma distribution with scale 1 and shape  $k$ , which was varied (see Table 2.3 for other values of this parameter used). To assess the impact of using binary (instead of continuous environments), a fraction  $f_B$  of the generated environments were binarised. Binarisation was achieved by converting the  $t$  most extreme continuous environmental values to ‘1’ and the remainder to ‘0’ where  $t$  was selected to give an event frequency  $v$ .  $f_B$  and  $v$  were varied whilst fixing the other parameter, using default settings  $f_B = 1$  and  $v = 0.02$  (see Table 2.3 for other values of the parameters used).

Further calibration experiments were conducted for these simulation settings, using 5,000 individuals and 60 environmental variables, considering the extreme cases where all environments were very skewed ( $k = 0.1$ , Fig. 2.7a, b) and all environments ( $f_B = 1$ ) were binary with rare event frequency ( $v = 0.005$ , Fig. 2.7c, d). In these settings, both the StructLMM association and interaction tests were assessed when no causal variants were simulated ( $v_g = 0$ , Fig. 2.7a, c) and StructLMM-int when persistent genetic effects were simulated ( $v_g = 0.006$ ,  $\rho = 0$ , Fig. 2.7b, d).

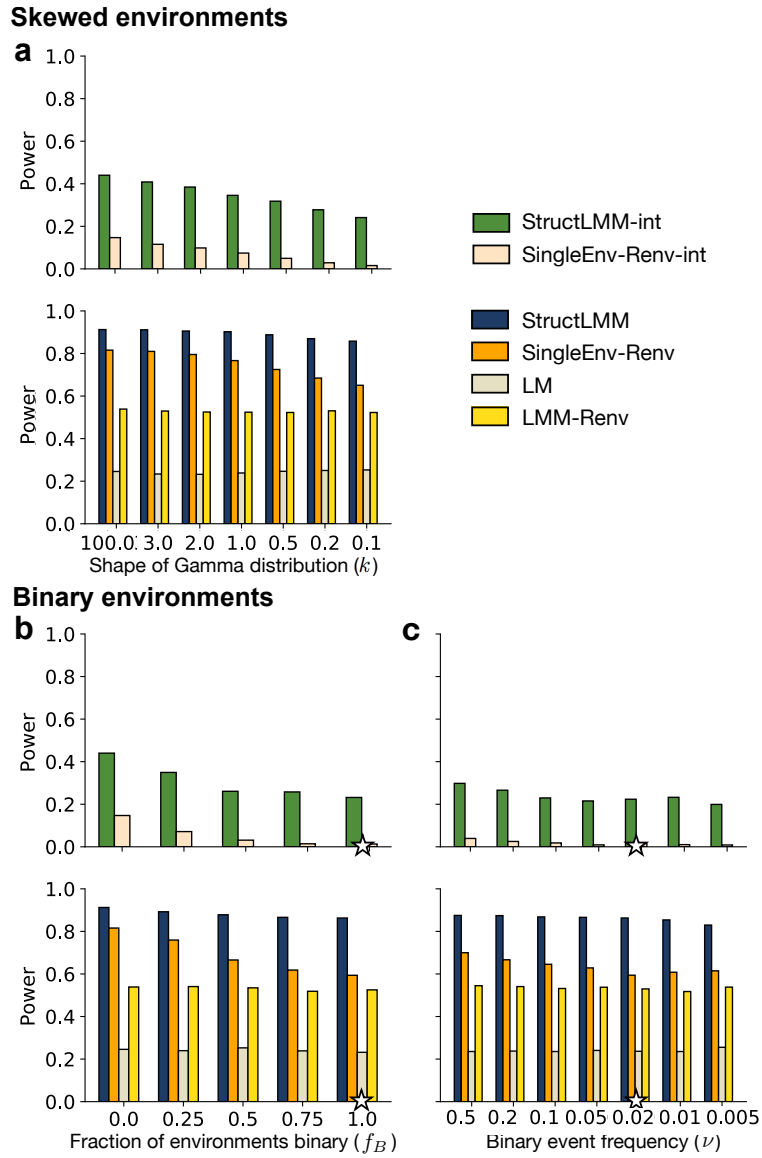


**Fig. 2.7 Calibration of StructLMM interaction and association tests when environments are simulated to be skewed or binary with rare event frequency** | QQ plots of negative log P values when all environments used to simulate the additive environmental effect (no simulated  $G \times E$  effects) are (a-b) highly skewed (shape parameter  $k = 0.1$ ), (c-d) are binary with rare event frequency ( $v = 0.005$ ). In panels a and c, QQ plots of negative log P values from StructLMM-int (green) and StructLMM (blue) when no genetic effects are simulated (no G, no  $G \times E$ ,  $v_g = 0$ ) and in panels b and d analogous QQ plots to assess the calibration of StructLMM-int when persistent genetic effects are simulated (no  $G \times E$ ,  $v_g = 0.006$ ,  $\rho = 0$ ). The genomic inflation factors  $\lambda_{GC}$  (denoted by  $\lambda$ ) for the interaction and association tests are displayed in the top left of each plot.

I then examined the effect of the environmental skew on the power of StructLMM compared to other methods. First the skew of the environmental variables was



increased by decreasing the shape parameter ( $k$ , whilst setting the scale parameter to 1) of the Gamma distribution from 100 (environment distribution is approximately Gaussian) to 0.1 (environments are highly skewed, Fig. 2.8a), noting that StructLMM increasingly outperforms other baseline methods that account for interaction effects as the environmental skew increases. Next, the fraction of environments that were binary ( $f_B$ ) was increased from 0 to 1 (Fig. 2.8b), which as expected, results in some loss of power due to loss of information but noting that StructLMM increasingly outperforms other methods that account for interaction effects, likely due to the fact that combining information from multiple environments retains more structure and thus information than assessing binary environments individually. The event frequency ( $v$ ) of these binary events was also decreased from 0.5 to 0.005 (Fig. 2.8c), which as expected gave similar results to increasing the skew of continuous environments.



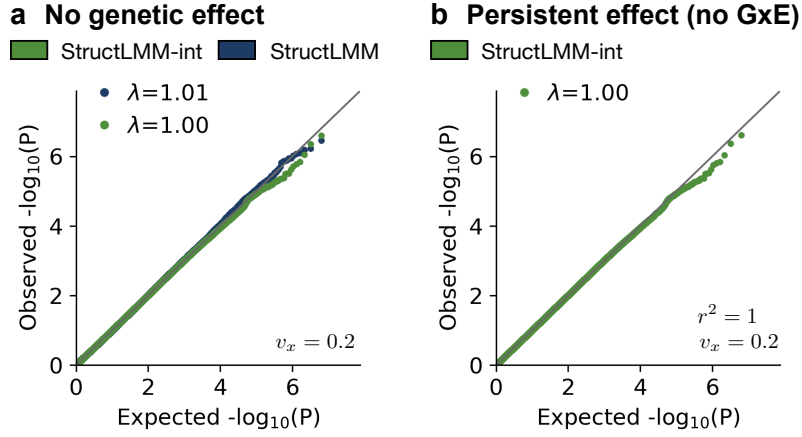
**Fig. 2.8 Power comparison of different methods examining the effect of using skewed and binary environments** | Power comparison of alternative methods for detecting interactions (top panels) and associations (bottom panels) based on simulated data. (a) Power comparison when varying the shape,  $k$ , of the Gamma distribution (lower shape values correspond to more skewed environments, shape 100 corresponds to approximately Gaussian environments). (b-c) Power comparison when simulating binary environments. (b) Power comparison varying the fraction of environments that are binary,  $f_B$  (with constant event frequency  $v = 0.02$ ). (c) Power comparison varying the event frequency of the binary event,  $v$  (when all environments are binary, i.e.  $f_B = 1$ ). In the top panels, considered are the StructLMM interaction test (StructLMM-int) and the default implementation of the single-environment interaction tests (SingleEnv-Renv-int, Table 2.2). In the bottom panels, considered are the StructLMM association test, the default 2 df fixed effect test that tests for associations whilst accounting for possible heterogeneity in the variant effect due to interactions with a single environment (SingleEnv-Renv, Table 2.2), and linear models that test for persistent effects (LM and LMM-Renv, Table 2.2). Models are assessed in terms of power (FWER < 1%) for detecting simulated causal variants. Stars denote default values of genetic parameters, which are retained when varying other parameters (see Table 2.3).

## Gene-environment correlations

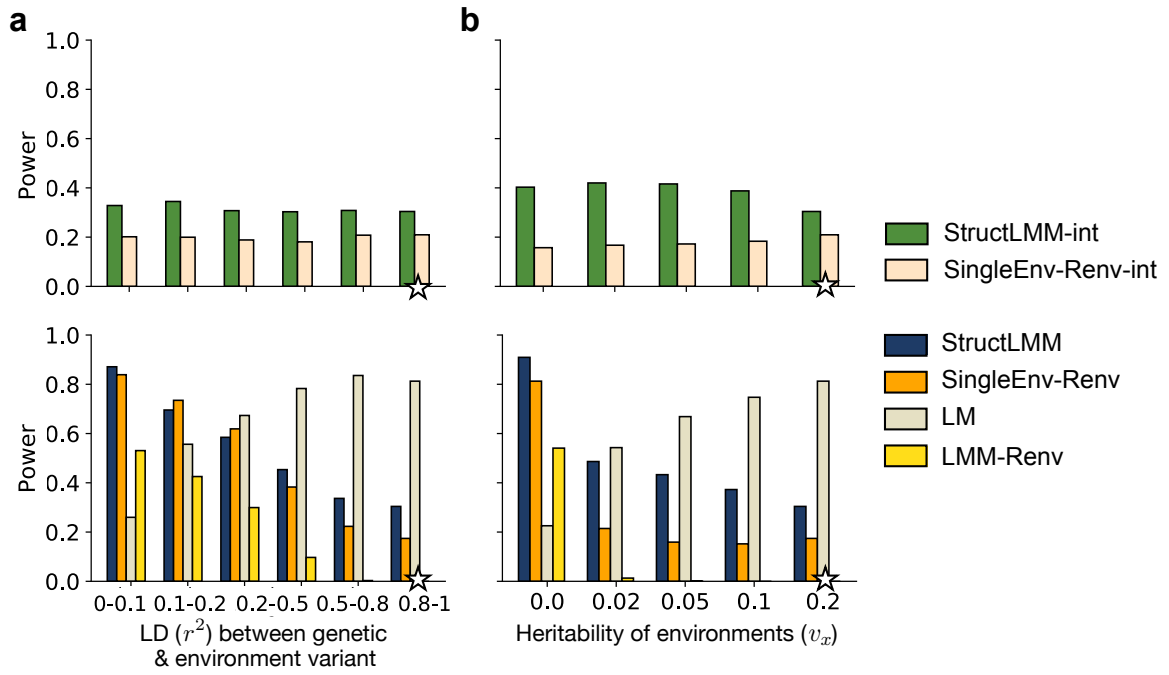
To generate heritable environments,  $\mathbf{C}$  was calculated directly from the randomly selected UK Biobank environment data, whilst  $\mathbf{R}$  was calculated as  $\mathbf{R} = v_x \mathbf{x}_e \mathbf{x}_e^T + (1 - v_x) \hat{\mathbf{R}}$ , where  $v_x$  is the fraction of the environmental variance driven by genetics,  $\mathbf{x}_e$  is the selected variant associated with the environments and  $\hat{\mathbf{R}}$  the sample covariance calculated directly from the UK Biobank environmental data.  $v_x$  and the LD,  $r^2$ , between  $\mathbf{x}_e$  and the selected variant  $\mathbf{x}$  used to simulate the  $\mathbf{g}$  and  $\mathbf{i}$  components of the phenotype (see Section 2.4.2) were varied whilst fixing the other parameter, using default settings  $v_x = 0.2$  and  $r^2 = 1$  (see Table 2.3 for other values of the parameters used).

Again calibration experiments were conducted, using 5,000 individuals and 60 environmental variables, considering the extreme case when all environments were heritable ( $v_x = 0.2$ ). In the setting where persistent genetic effects were simulated, the variant associated with the environments was the same variant that drives the persistent effect ( $r^2 = 1$ , Fig. 2.9b). Both the StructLMM association and interaction tests were assessed when no causal variants were simulated ( $v_g = 0$ , Fig. 2.9a) and StructLMM-int when persistent genetic effects were simulated ( $v_g = 0.006$ ,  $\rho = 0$ , Fig. 2.9b). Both tests were always calibrated, demonstrating that StructLMM can be robustly applied to different settings.

The effect on power of heritable environments was then examined. First, the LD ( $r^2$ ) between the variant associated with the environments and the variant used to simulate the genetic effect ( $\mathbf{g}$  and  $\mathbf{i}$ , see Section 2.4.2) was increased, considering LD bins and the setting where this was the same variant (Fig. 2.10a). Second, the average heritability ( $v_x$ ) of the environments was varied between 0.0 and 0.2 (Fig. 2.10b). Whilst the power of the interaction tests are largely unaffected by the degree of heritability and LD between variants, all association tests (with the exception of the LM) lose power as the heritability and LD increases. This is because it is not possible to distinguish between the two causal mechanisms, (i) direct effect of the variant on the phenotype and (ii) indirect effect of the variant on the phenotype where the environments act as an intermediate. Hence, when additive environment effects are not accounted for (as in the LM), there is no loss in power. This highlights that for association testing, the choice of null model can have a large impact on the power to discover associations.



**Fig. 2.9 Calibration of StructLMM in the presence of heritable environments** | QQ plots of negative log P values when all environments used to simulate the additive environmental effect (no simulated  $G \times E$  effects) are heritable ( $v_x = 0.2$ ). In panel **a**, QQ plots of negative log P values from StructLMM-int (green) and StructLMM (blue) when no genetic effects are simulated (no G, no  $G \times E$ ,  $v_g = 0$ ) and in panel **b**, analogous QQ plots to assess the calibration of StructLMM-int when persistent genetic effects are simulated (no  $G \times E$ ,  $v_g = 0.006$ ,  $\rho = 0$ ), where the same variant is used to simulate the heritable environment and the persistent genetic effect ( $r^2 = 1$ ). The genomic inflation factors  $\lambda_{GC}$  (denoted by  $\lambda$ ) for the interaction and association tests are displayed in the top left of each plot.



**Fig. 2.10 Power comparison of different methods in the presence of heritable environments** | Power comparison of alternative methods for detecting interactions (top panels) and associations (bottom panels) based on simulated data, increasing (a) the LD ( $r^2$ ) between the variant associated with the environments and the variant driving the G and G×E components of the phenotype and (b) the average fraction of the environments variance explained by genetic effects ( $v_x$ , heritability). In the top panels, considered are the StructLMM interaction test (StructLMM-int) and the default implementation of the single-environment interaction test (SingleEnv-Renv-int, Table 2.2). In the bottom panels, considered are the StructLMM association test, the default 2 df fixed effect test that tests for associations whilst accounting for possible heterogeneity in the variant effect due to interactions with a single environment (SingleEnv-Renv, Table 2.2), and linear models that test for persistent effects (LM and LMM-Renv, Table 2.2). Models are assessed in terms of power (FWER < 1%) for detecting simulated causal variants. Stars denote default values of genetic parameters, which are retained when varying other parameters (see Table 2.3).

Together, the results for general simulation settings (described in Section 2.4.4) and settings in which environmental variables with specific properties are used (described in this section), demonstrate that StructLMM is robust. It retains calibration across all considered scenarios, notably including the scenario in which the environments themselves are driven by the same variant as that driving the persistent and G×E effect. In comparison, other implementations of multi-environment interaction and association tests are not always calibrated, highlighting the benefit of using a random effect term to aggregate across environments, such that the number of degrees of freedom of StructLMM is independent of the number of considered environments.

In addition, StructLMM offers power advantages over other baseline methods, in particular when the number of environments contributing to the  $G \times E$  effect is large, when the environments are non-Gaussian and the fraction of the genetic variance driven by interaction effects ( $\rho$ ) is moderate to high, whilst maintaining similar power to linear models when no  $G \times E$  effects are present. StructLMM will have marginally less power than linear models when no  $G \times E$  effects are present, since it is penalised for grid searching over different values of  $\rho$ . As a result it is not designed to be a replacement for linear models but an alternative to identify additional loci, in particular those with moderate to strong  $G \times E$  effects.

## 2.5 Summary and discussion

In this chapter, I have shown that a random effect design can be used to test for  $G \times E$  effects jointly at multiple environmental variables using a single covariance term (i.e. only 1 df is required to account for  $G \times E$  at multiple environmental variables). This framework can be used to perform both an interaction test and an association test that accounts for possible heterogeneity in variant effects due to differences in environmental exposures, both of which rely on a score-based testing procedure. The association test is a new statistical test, differing from existing set tests, for which I explicitly derive the mathematics that can be used to evaluate the significance of the test statistics and thus obtain P values (see Section 2.2.5). I note that SKAT-O<sup>123</sup> is a special case of the new statistical test derived in this chapter when the symmetric matrix,  $\Sigma$ , is replaced by the identity matrix  $\mathbf{I}$ . As described in Section 2.2.6, this method has been implemented in a computationally efficient manner such that it scales linearly with the number of samples, rather than cubically as would be the case if implemented naively. The method is broadly applicable, allowing for both interaction testing with hundreds of environments and the identification of epistatic interaction effects, using cohorts containing hundreds of thousands of samples. These methods are freely available at <https://github.com/limix/struct-lmm> with tutorials and illustrations on how to use the model available at <http://struct-lmm.readthedocs.io>.

Through the use of extensive simulation studies, I have shown that StructLMM is calibrated across all settings considered and enjoys power gains across a broad range of settings. It should be noted that this association test will be marginally less powered than linear models when no  $G \times E$  effects exist, since there is some penalisation for grid searching over different values of  $\rho$  and as a result StructLMM

is not designed to replace the linear model but rather to act as an alternative tool to identify additional loci, in particular those with moderate to strong  $G \times E$  effects.

Currently, StructLMM is applicable to quantitative traits but the test can be readily adapted for application to binary traits. In fact the SKAT-O method<sup>123</sup> was originally formulated based on application to binary traits. A potential limitation is the speed of the method as generalised linear mixed model parameter inference is slower than LMM parameter inference and whilst StructLMM is implemented in a computationally efficient manner it is already slower than a standard LMM.

Throughout this work, only linear covariance functions were considered but in principle any covariance function can be considered (see Rasmussen *et al.*<sup>256</sup> for descriptions of other covariance functions) and this may be particularly pertinent as sample sizes increase such that there is sufficient power to detect non-linear interactions.

Another limitation that is harder to address, is that currently the method can only handle a single low rank random effect term, which is used to account for the additive environment due to correlations between the environments and the phenotypes. In theory, the method could be extended to handle a second random effect term that can account for confounding due to population structure, thus allowing the inclusion of related individuals. However, the kinship matrix is in general full rank; thus in order to retain computational efficiency, particularly important for application to cohorts of large sample size, this matrix would need to be built using suitable low-rank approximations (e.g. as described by Listgarten *et al.*<sup>274</sup>).

Application of the StructLMM method, proposed in this chapter, to a real dataset will be the focus of the next chapter.





# Chapter 3

## Application of StructLMM to identify genotype-environment interaction effects that influence body mass index in UK Biobank

### 3.1 Introduction

Body mass index (BMI), defined as  $\frac{\text{Weight (kg)}}{(\text{Height (m)})^2}$ , is a surrogate measure of adiposity or overall obesity<sup>275</sup>, with guidelines available from the World Health Organisation (WHO) for classifying individuals as underweight, healthy, overweight or obese (which can be further subcategorised)<sup>276</sup>. Other commonly used measures to assess adiposity and in particular, body fat distribution include waist circumference (WC) and waist-to-hip ratio (WHR), both of which can be adjusted for BMI. Alternatively, body composition can be assessed through the use of a dual energy X-ray absorptionmetry (DEXA) scan<sup>276–278</sup>.

Obesity (characterised by an excess of adipose tissue) and overweight prevalence has more than doubled since 1970 and if the rate of increase (as of 2005) continues, it is estimated that by 2030, 38% of the world’s adult population will be overweight and an additional 20% obese<sup>279–284</sup>. Consequently, due to the associated increase in risk of other diseases, it is one of the largest global health burdens<sup>277</sup>. Comorbidities, include type 2 diabetes, cardiovascular disease, stroke, some types of cancer<sup>275,281,285–291</sup> and in addition, maternal obesity can lead to congenital abnormalities<sup>292</sup>. There are also less severe associated health outcomes, including

osteoarthritis, infertility, asthma and sleep apnoea<sup>285,293,294</sup>. Thus a good understanding of the factors governing BMI is important.

Whilst it is widely accepted that changes in the global food system, resulting in increased availability and a shift to convenience based diets, coupled with an increase in sedentary lifestyles are major contributors to this global epidemic<sup>277,295–300</sup>, there is also a strong genetic contribution to BMI risk. Heritability estimates range from 30 – 70%<sup>49,301–304</sup>.

Early genetic studies focussed on monogenic forms of obesity, where mutations in a single gene, for example, *LEP*, *LEPR*, *POMC* and *MC4R*, drive the observed phenotype<sup>305–312</sup>. However, for the majority of individuals, BMI is a complex trait and the advent of GWAS, in 2005, enabled exploration of the polygenic factors responsible<sup>295</sup>.

The first locus identified through this approach in 2007, that was later replicated in an independent study, lies within the *FTO* gene<sup>277,313</sup>. To date, this is the locus known to have the largest effect on BMI in Europeans (considering only common variants)<sup>314</sup>, although it is estimated to explain only  $\approx 0.34\%$  of BMI variance<sup>277,297</sup>. This was followed a year later by the identification of a signal 188 kb downstream of *MC4R*<sup>277,315</sup>, a gene already associated with monogenic forms of obesity. A number of further loci have since been identified, with the GIANT consortium being a major driver of these discoveries<sup>296,297,316</sup>; the 2015 meta-analysis, based on 339,224 individuals identified 97 loci, 56 of which were novel<sup>316</sup>. Despite the large number of loci identified, these were estimated to explain only  $\approx 2.7\%$  of BMI variation. In an attempt to identify more loci associated with BMI risk, a very large cohort of 681,275 European individuals was recently analysed, obtained by combining the samples collated by the GIANT consortium and UK Biobank<sup>271</sup>. This resulted in the identification of 536 loci of which 484 had not been previously detected<sup>317</sup>.

The rapid change in lifestyles over the past few decades, during which time the genetic pool has stayed relatively constant, has coincided with a substantial density shift towards the upper end of the BMI population distribution<sup>298,312</sup>. This has fostered an interest in studying G $\times$ E effects on BMI<sup>298,312</sup>. Several studies provide evidence to suggest that genetically predisposed individuals are at a greater risk of BMI increase in obesogenic environments<sup>318,319</sup>. The first examined 907 non-Hispanic White adults from the Fels Longitudinal Study, with individuals binned into five groups of approximately equal size according to year of birth (1939 or earlier, 1940 – 1949, 1950 – 1959, 1960 – 1969 and 1970 or later)<sup>318</sup>. Year of birth was subsequently used as a proxy for an individuals environmental exposure. A significant interaction

effect ( $P < 0.001$ ) on BMI between year of birth and an unweighted GRS (built using the 32 SNPs associated with BMI reported by Speliotes *et al.*<sup>297</sup> and noting that there was no mean difference in GRS according to year of birth), whilst appropriately accounting for other factors such as age and sex was identified<sup>318</sup>. Similarly, the second study which uses longitudinal data from the Offspring Cohort of the Framingham Heart Study partitioned individuals based on year of birth into those born pre and post 1942. A significant interaction effect ( $P < 0.05$ ) between year of birth and a variant in the *FTO* gene (*rs9939609*) on BMI was identified, again accounting for appropriate factors as covariates<sup>319</sup>.

Genotype-environment interaction studies have resulted in the identification of significant G×E effects on BMI with a range of environments, including physical activity<sup>196–199,206,212,213,320–324</sup>, dietary components<sup>196,200–205,214,325</sup>, alcohol consumption<sup>320</sup>, smoking status<sup>196,320</sup>, socioeconomic status (often defined by the Townsend deprivation index (TDI) in the UK)<sup>320,326</sup>, mental health<sup>320</sup>, sleep patterns<sup>320</sup>, gender<sup>327–329</sup> and age<sup>330,331</sup>. Genotype-age interaction effects can be difficult to interpret as they can reflect either an accumulation of non-specific environmental exposures or that biological mechanisms alter with age<sup>312,330</sup>. Some of these studies have tested for G×E effects between individual environments and single genetic variants, repeatedly finding significant interaction effects at the *FTO* locus (see Section 2.2.7 for method details)<sup>197–206</sup>. Other studies have used a set test based approach, with as many as 94 variants, building unweighted<sup>206,212,213,321,322,330</sup> or weighted GRS, testing for interaction effects between this score and individual environmental variables<sup>214,320,323,325,326</sup> (see Section 2.2.7 for method details).

However, many of these environments are correlated and as a result, some of these G×E effects may not be independent of one another but instead tag the same environmental driving factor, which is potentially unobserved and/or directly unmeasurable. Hence, claims that the single environmental exposure examined in an interaction study, is *the* one responsible for the observed interaction effect and should be *the focus* of public health policies, are likely overstated<sup>197,207,214</sup>. A recent study by Young *et al.*<sup>195</sup>, largely overcame these issues by jointly testing for interaction effects with multiple environments (see Section 2.2.7 for further details). Similarly, the developed StructLMM method, described in Chapter 2, also mitigates these issues by testing for interaction effects with multiple environments. In addition, subsequent exploration of the putative driving environments using StructLMM is based on a backward elimination procedure, thereby negating the issue of correlated environments (see Section 3.2.2). Furthermore, I demonstrate that StructLMM can be used to identify individuals within the population that are at increased (or

decreased) trait and disease risk based on their aggregate environmental profiles (and genotypes), which is perhaps more pertinent than trying to identify specific environmental drivers.

The previous study<sup>195</sup> that accounts for multiple environments when testing for  $G \times E$  effects on BMI, uses the UK Biobank cohort<sup>271</sup>. UK Biobank is a large prospective cohort study, comprising of  $\sim 500,000$  British individuals, aged between 40 and 69 years at recruitment. In-depth phenotype and environment data is available for these individuals, making this an ideal cohort for multi-environment interaction studies (I refer the reader to Sudlow *et al.*<sup>271</sup> and Bycroft *et al.*<sup>215</sup> and <http://www.ukbiobank.ac.uk> for further details).

Given the evidence for  $G \times E$  effects to alter BMI risk, potentially due to multiple environments that are not independent of another and the availability of such environments in a single cohort of individuals (UK Biobank), this is a good setting to demonstrate the practical utility of StructLMM.

In this chapter, I will use 64 lifestyle based factors derived from the available UK Biobank data to test for interactions and associations and then further explore some of the findings. Specifically, in Section 3.2, I will provide details of methods implemented as part of StructLMM that can be used for exploring significant findings. In Section 3.3, I will describe general methods that are used throughout this chapter, including data pre-processing steps and in Section 3.4, I will outline the results of this application.

The material presented in this chapter has been published by Nature Genetics<sup>245</sup>. A copy of this publication can be found in Appendix A (apart from the Supplementary Tables which are available at <https://www.nature.com/articles/s41588-018-0271-0>). The implementation of methods used to explore identified loci (described in Section 3.2) was joint work with Francesco Paolo Casale.

## 3.2 Methods to explore identified loci

In this section, I will describe the methods that have been implemented as part of StructLMM, that can be used to explore significant interaction or association variants. This includes, estimating the fraction of the genetic variance that is driven by  $G \times E$ , exploration of the environments that drive the interaction effects and identification of environmental profiles and thus individuals (if they carry a risk increasing allele) that are at increased or decreased disease or trait risk. Finally, the

computational complexity of these downstream interpretation methods is discussed.

### 3.2.1 Estimating the fraction of the genetic variance driven by $\mathbf{G} \times \mathbf{E}$

StructLMM can be used to estimate the fraction of phenotypic variance driven by marginal genetic effects ( $\mathbf{G}$ ), interaction effects ( $\mathbf{G} \times \mathbf{E}$ ) and if desired marginal environment effects ( $\mathbf{E}$ ), denoted by  $\text{var}^{(\mathbf{G})}$ ,  $\text{var}^{(\mathbf{G} \times \mathbf{E})}$  and  $\text{var}^{(\mathbf{E})}$ , respectively. This is achieved using the interaction test marginal distribution (described by Eq. 2.27, Chapter 2) for parameter inference, obtaining the MLE  $\hat{\boldsymbol{\alpha}}$ ,  $\hat{\boldsymbol{\beta}}_{\mathbf{G}}$ ,  $\hat{\sigma}_{\mathbf{G} \times \mathbf{E}}^2$ ,  $\hat{\sigma}_e^2$  and  $\hat{\sigma}_n^2$ . The variance explained by the different components is then estimated as:

$$\begin{aligned} \text{var}^{(\mathbf{G})} &= \text{var}_s(\mathbf{x} \hat{\boldsymbol{\beta}}_{\mathbf{G}}) \\ &= \hat{\beta}_{\mathbf{G}}^2 \text{var}_s(\mathbf{x}), \end{aligned} \quad (3.1)$$

$$\begin{aligned} \text{var}^{(\mathbf{G} \times \mathbf{E})} &= \mathbb{E}[\text{var}_s(\mathbf{x} \odot \boldsymbol{\beta}_{\mathbf{G} \times \mathbf{E}})] \\ &= \text{var}_M(\hat{\sigma}_{\mathbf{G} \times \mathbf{E}}^2 \text{diag}(\mathbf{x}) \boldsymbol{\Sigma} \text{diag}(\mathbf{x})) \\ &= \text{var}_M(\hat{\sigma}_{\mathbf{G} \times \mathbf{E}}^2 (\mathbf{x} \odot \mathbf{E})(\mathbf{x} \odot \mathbf{E})^T), \end{aligned} \quad (3.2)$$

$$\begin{aligned} \text{var}^{(\mathbf{E})} &= \mathbb{E}[\text{var}_s(\mathbf{u})] \\ &= \text{var}_M(\hat{\sigma}_e^2 \boldsymbol{\Sigma}) \\ &= \text{var}_M(\hat{\sigma}_e^2 \mathbf{E} \mathbf{E}^T), \end{aligned} \quad (3.3)$$

where  $\text{var}_s$  denotes the sample variance and  $\text{var}_M(\mathbf{K})$  denotes the expected sample variance of  $\mathbf{z}$ , an  $N \times 1$  vector following  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ , where  $\mathbf{K}$  is an  $N \times N$  matrix.

Using  $\text{var}_M(\mathbf{K}) = \frac{1}{N-1} \text{tr}(\mathbf{P} \mathbf{K})$ , where  $\mathbf{P} = \mathbf{I}_N - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T$ <sup>257,332</sup>, it follows that:

$$\text{var}^{(\mathbf{G} \times \mathbf{E})} = \frac{\hat{\sigma}_{\mathbf{G} \times \mathbf{E}}^2}{N-1} \text{tr}(\mathbf{P} (\mathbf{x} \odot \mathbf{E})(\mathbf{x} \odot \mathbf{E})^T), \quad (3.4)$$

$$\text{var}^{(\mathbf{E})} = \frac{\hat{\sigma}_e^2}{N-1} \text{tr}(\mathbf{P} \mathbf{E} \mathbf{E}^T). \quad (3.5)$$

Noting that  $\frac{1}{N} \mathbf{1}_N \mathbf{1}_N^T$  is a projection matrix (spanning the space  $\mathbf{1}_N$ ), such that

$\mathbf{P}\mathbf{P} = \mathbf{P}$  and  $\mathbf{P} = \mathbf{P}^T$  and  $\text{tr}(\mathbf{A}\mathbf{B}) = \text{tr}(\mathbf{B}\mathbf{A})$ , then:

$$\begin{aligned}
\text{var}^{(\text{G} \times \text{E})} &= \frac{\hat{\sigma}_{\text{G} \times \text{E}}^2}{N-1} \text{tr}(\mathbf{P}(\mathbf{x} \odot \mathbf{E})(\mathbf{x} \odot \mathbf{E})^T) \\
&= \frac{\hat{\sigma}_{\text{G} \times \text{E}}^2}{N-1} \text{tr}(\mathbf{P}\mathbf{P}(\mathbf{x} \odot \mathbf{E})(\mathbf{x} \odot \mathbf{E})^T) \\
&= \frac{\hat{\sigma}_{\text{G} \times \text{E}}^2}{N-1} \text{tr}(\mathbf{P}(\mathbf{x} \odot \mathbf{E})(\mathbf{x} \odot \mathbf{E})^T \mathbf{P}) \\
&= \frac{\hat{\sigma}_{\text{G} \times \text{E}}^2}{N-1} \text{tr}([\mathbf{P}(\mathbf{x} \odot \mathbf{E})][\mathbf{P}(\mathbf{x} \odot \mathbf{E})]^T) \\
&= \frac{\hat{\sigma}_{\text{G} \times \text{E}}^2}{N-1} \|\mathbf{P}(\mathbf{x} \odot \mathbf{E})\|_{\text{F}}^2,
\end{aligned} \tag{3.6}$$

where  $\|\cdot\|_{\text{F}}$  is the Frobenius norm. Similarly,

$$\begin{aligned}
\text{var}^{(\text{E})} &= \frac{\hat{\sigma}_e^2}{N-1} \text{tr}(\mathbf{P}\mathbf{E}\mathbf{E}^T) \\
&= \frac{\hat{\sigma}_e^2}{N-1} \|\mathbf{P}\mathbf{E}\|_{\text{F}}^2.
\end{aligned} \tag{3.7}$$

Therefore the estimated fraction of genetic variance explained by  $\text{G} \times \text{E}$  effects,  $\rho$ , is:

$$\rho = \frac{\text{var}^{(\text{G} \times \text{E})}}{\text{var}^{(\text{G} \times \text{E})} + \text{var}^{(\text{G})}}. \tag{3.8}$$

Whilst it can be seen that the parameter  $\rho$  described here takes a similar form to the grid search parameter  $\rho$  used in the StructLMM joint association test (see Section 2.2.2), this estimate is based on the MLE and thus is not constrained to a grid of predefined values. As I have already explained in Section 2.2.2, performing a joint association test in the same vein as described here (i.e. a 2 df joint association test), is not possible as there is no closed form solution for obtaining P values when jointly testing two variance component parameters that lie on the boundary of the parameter space under the null. Using the estimate of  $\rho$  obtained here to calculate  $Q_\rho$  and the corresponding P value (see Section 2.2.5) is also not valid since the tests are not independent.

### 3.2.2 Exploration of the environments that drive the $\text{G} \times \text{E}$ effects

StructLMM can be used to explore the environments that drive the observed  $\text{G} \times \text{E}$  effects, through the use of Bayes factors, which quantify the support for different compared models. Specifically, this is achieved by subtracting the log marginal

likelihood of a model with environments excluded from that of a model containing all environments to give a  $\log(\text{Bayes factor})$ . By calculating the effect of excluding environments, rather than including them, I account for other correlated environments and thus gain a clearer understanding of the environments have the most bearing on the observed interaction effects.

Explicitly, denoting the full set of  $L$  environments,  $\varepsilon = \{e_1, e_2, \dots, e_L\}$ , the evidence for a subset of these environments,  $\varepsilon_i = \{e_1, e_2, \dots, e_{L_i}\}$  (where  $\varepsilon_i \subseteq \varepsilon$  with  $|\varepsilon_i| = L_i$ ), driving the observed  $G \times E$  effect at a given variant is given by:

$$\log(\text{Bayes factor})(\varepsilon_i) = \text{LML}(M_\varepsilon) - \text{LML}(M_{\varepsilon \setminus \varepsilon_i}). \quad (3.9)$$

$\setminus$  denotes the set difference, such that the  $\text{LML}(M_\varepsilon)$  and  $\text{LML}(M_{\varepsilon \setminus \varepsilon_i})$  denote the log marginal likelihoods of the models containing all  $L$  environments or the reduced set of environments in the  $G \times E$  term of the model described by Eq. 2.27 (see Chapter 2), respectively. Note that for both models, all  $L$  environments are included in the additive environment term (E), such that the only difference between the two considered models is in the  $G \times E$  term.

I consider removing each of the  $L$  environments individually, providing putative evidence of whether each environment is likely or unlikely to contribute to the  $G \times E$  effect, where a  $\log(\text{Bayes factor}) > 0$  and  $\log(\text{Bayes factor}) < 0$  are evidence of contribution and no contribution, respectively. The strength of this evidence can be further classified based on the log Kass Rafferty scale<sup>333</sup>, where  $|\log(\text{Bayes factor})| \leq 1$  is ‘not worth more than a bare mention’,  $1 < |\log(\text{Bayes factor})| \leq 3$  is ‘positive’ and  $|\log(\text{Bayes factor})| > 3$  is ‘strong’.

In addition, I can identify a putative causal set of driving environments using a greedy backward elimination procedure. Specifically, I initially identify the environment with the most evidence for driving the observed  $G \times E$  effect, based on the results from removing each environment individually in turn. This environment is then always removed from the model (i.e. always included in  $\varepsilon_i$ ) and then a Bayes factor, excluding in addition each of the  $L - 1$  remaining environments one by one (i.e.  $L_i = 2$ ) is calculated. I then assess which of these additional  $L - 1$  environments provides the most evidence for driving the observed  $G \times E$  effect and this environment is then permanently removed from the model (i.e. also included in  $\varepsilon_i$ ). This process is iteratively repeated, stopping when there is positive evidence based on the log Kass Rafferty scale<sup>333</sup> that I have selected a full set of environments that can explain the observed  $G \times E$  effect. Explicitly, this occurs at the first time when  $\text{LML}(M_{\varepsilon \setminus \varepsilon_i}) - \text{LML}(M_0) < 1$ , where  $M_0$  is the model described by Eq. 2.27

(see Chapter 2) with no modelled  $G \times E$  term. I also examine the set of environments required to obtain strong evidence<sup>333</sup> that I have selected a set of environments that explain the observed  $G \times E$  effect (i. e. the first time when  $\text{LML}(M_{\varepsilon \setminus \varepsilon_i}) - \text{LML}(M_0) < 3$ ).

### 3.2.3 Estimation of per-individual allelic effect sizes due to $G \times E$

StructLMM can be used to estimate per-individual allelic effect sizes based on the environmental profiles present in a population, thus enabling identification of individuals at increased or decreased trait risk if they carry the alternative allele. This is achieved by making out-of-sample predictions for the total genetic component ( $G + G \times E$ ) of the phenotype using best linear unbiased predictors (BLUP)<sup>334–336</sup>. I note that the BLUP is equivalent to the mean of the conditional (conditioning on the observed phenotypes  $\mathbf{y}$ ) multivariate Gaussian distribution<sup>337</sup>.

Starting from the interaction test marginal distribution described by Eq. 2.27 (see Chapter 2)<sup>†</sup> to obtain MLE  $\hat{\boldsymbol{\alpha}}$ ,  $\hat{\beta}_G$ ,  $\hat{\sigma}_{G \times E}^2$ ,  $\hat{\sigma}_e^2$  and  $\hat{\sigma}_n^2$ , it follows that the MLE of the covariance (as described in Eq. 2.27),  $\hat{\mathbf{K}}$  can be written as:

$$\hat{\mathbf{K}} = \hat{\sigma}_{G \times E}^2 (\mathbf{x} \odot \mathbf{E})(\mathbf{x} \odot \mathbf{E})^T + \hat{\sigma}_e^2 \boldsymbol{\Sigma} + \hat{\sigma}_n^2 \mathbf{I}_N. \quad (3.10)$$

Denoting the total predicted genetic component of the phenotype as  $\mathbf{y}_\star$ , which is an  $N_\star \times 1$  vector, where  $N_\star$  is the number of individuals for which predictions are to be made and letting  $\mathbf{x}_\star$  be the  $N_\star \times 1$  genotype dosage vector for the  $N_\star$  individuals and  $\mathbf{E}_\star$  be the  $N_\star \times L$  environmental matrix describing the  $L$  environmental variables for the  $N_\star$  individuals, then:

$$\mathbb{E}[\mathbf{y}_\star | \mathbf{y}] = \underbrace{\mathbf{x}_\star \hat{\beta}_G}_{G_\star} + \underbrace{\hat{\sigma}_{G \times E}^2 (\mathbf{x}_\star \odot \mathbf{E}_\star)(\mathbf{x} \odot \mathbf{E})^T \hat{\mathbf{K}}^{-1} (\mathbf{y} - \mathbf{W} \hat{\boldsymbol{\alpha}} - \mathbf{x} \hat{\beta}_G)}_{(G \times E)_\star}. \quad (3.11)$$

I can use the same set of  $N_\star$  individuals to:

1. Predict the genetic component of the phenotype using the reference allele,  $x^{(r)}$ , such that the  $N_\star \times 1$  genotype dosage vector is  $\mathbf{x}_\star^{(r)} = [x^{(r)}, x^{(r)}, \dots, x^{(r)}]^T$ , then:

$$\mathbf{y}_\star^{(r)} = \mathbf{x}_\star^{(r)} \hat{\beta}_G + \hat{\sigma}_{G \times E}^2 (\mathbf{x}_\star^{(r)} \odot \mathbf{E}_\star)(\mathbf{x} \odot \mathbf{E})^T \hat{\mathbf{K}}^{-1} (\mathbf{y} - \mathbf{W} \hat{\boldsymbol{\alpha}} - \mathbf{x} \hat{\beta}_G). \quad (3.12)$$

---

<sup>†</sup>using that  $\text{diag}(\mathbf{x}) \boldsymbol{\Sigma} \text{diag}(\mathbf{x}) = (\mathbf{x} \odot \mathbf{E})(\mathbf{x} \odot \mathbf{E})^T$



2. Predict the genetic component of the phenotype using the alternate allele,  $x^{(a)}$ , such that the  $N_\star \times 1$  genotype dosage vector is  $\mathbf{x}_\star^{(a)} = [x^{(a)}, x^{(a)}, \dots, x^{(a)}]^T$ , then:

$$\mathbf{y}_\star^{(a)} = \mathbf{x}_\star^{(a)} \hat{\beta}_G + \hat{\sigma}_{G \times E}^2 (\mathbf{x}_\star^{(a)} \odot \mathbf{E}_\star) (\mathbf{x} \odot \mathbf{E})^T \hat{\mathbf{K}}^{-1} (\mathbf{y} - \mathbf{W} \hat{\boldsymbol{\alpha}} - \mathbf{x} \hat{\beta}_G). \quad (3.13)$$

Consequently, the predicted allelic effect,  $\beta_\star = \mathbf{y}_\star^{(a)} - \mathbf{y}_\star^{(r)}$ , that accounts for both G (persistent) and G×E (interaction) effects can be written as:

$$\beta_\star = (\mathbf{x}_\star^{(a)} - \mathbf{x}_\star^{(r)}) \hat{\beta}_G + \hat{\sigma}_{G \times E}^2 (\mathbf{x}_\star^{(a)} - \mathbf{x}_\star^{(r)}) \odot \mathbf{E}_\star (\mathbf{x} \odot \mathbf{E})^T \hat{\mathbf{K}}^{-1} (\mathbf{y} - \mathbf{W} \hat{\boldsymbol{\alpha}} - \mathbf{x} \hat{\beta}_G). \quad (3.14)$$

If the genotype dosages are encoded as 0, 1, 2, where 0 is homozygous reference, 1 is heterozygous and 2 homozygous alternative then,  $(\mathbf{x}_\star^{(a)} - \mathbf{x}_\star^{(r)}) = 1$  and the predicted allelic effect given by Eq. 3.14 is simply:

$$\beta_\star = \hat{\beta}_G + \hat{\sigma}_{G \times E}^2 \mathbf{E}_\star (\mathbf{x} \odot \mathbf{E})^T \hat{\mathbf{K}}^{-1} (\mathbf{y} - \mathbf{W} \hat{\boldsymbol{\alpha}} - \mathbf{x} \hat{\beta}_G). \quad (3.15)$$

If the genotype dosages are mean centred and standardised, such that  $x^{(r)} = \frac{-2p}{\sqrt{2p(1-p)}}$  and  $x^{(a)} = \frac{1-2p}{\sqrt{2p(1-p)}}$ , where  $p$  is the observed frequency of the minor allele, then  $(\mathbf{x}_\star^{(a)} - \mathbf{x}_\star^{(r)}) = \frac{1}{\sqrt{2p(1-p)}}$  and the allelic effect is given by:

$$\beta_\star = \frac{1}{\sqrt{2p(1-p)}} \hat{\beta}_G + \frac{1}{\sqrt{2p(1-p)}} \hat{\sigma}_{G \times E}^2 \mathbf{E}_\star (\mathbf{x} \odot \mathbf{E})^T \hat{\mathbf{K}}^{-1} (\mathbf{y} - \mathbf{W} \hat{\boldsymbol{\alpha}} - \mathbf{x} \hat{\beta}_G). \quad (3.16)$$

By setting  $\mathbf{E}_\star = \mathbf{E}$ , I can perform in-sample estimation of the allelic effects. Alternatively, if I randomly select a set of individuals from the original  $N$  to obtain the MLE and then set  $\mathbf{E}_\star = \mathbf{E}_{\text{remaining}}$ , where  $\mathbf{E}_{\text{remaining}}$  is the environment matrix of the remaining individuals, I can make out-of-sample predictions of the allelic effects.

### 3.2.4 Estimation of the aggregate environment driving the G×E effect at a variant

StructLMM can also be used to estimate the aggregate environment driving the G×E effect at a variant. As already noted in Sections 2.2.4 and 2.3.2, StructLMM can be derived from a multivariate model, in which as many fixed effects as there are environments are used to define the model. This model, described in Eq. 2.24,

can be rewritten as:

$$\begin{aligned}
\mathbf{y} &= \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{covariates}} + \underbrace{\mathbf{x}\beta_G}_G + \underbrace{\sum_{l=1}^L (\mathbf{x} \odot \mathbf{e}_l) \beta_{G \times E_l}}_{G \times E} + \underbrace{\sum_{l=1}^L \mathbf{e}_l \beta_{E_l}}_E + \underbrace{\boldsymbol{\epsilon}}_{\text{noise}} \\
&= \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{covariates}} + \underbrace{\mathbf{x}\beta_G}_G + \underbrace{\mathbf{x} \odot \sum_{l=1}^L \mathbf{e}_l \beta_{G \times E_l}}_{G \times E} + \underbrace{\sum_{l=1}^L \mathbf{e}_l \beta_{E_l}}_E + \underbrace{\boldsymbol{\epsilon}}_{\text{noise}} \\
&\quad \text{Aggregate environment} \\
&= \underbrace{\mathbf{W}\boldsymbol{\alpha}}_{\text{covariates}} + \underbrace{\mathbf{x}\beta_G}_G + \underbrace{\mathbf{x} \odot (\mathbf{E}\boldsymbol{\beta}'_{G \times E})}_{G \times E} + \underbrace{\sum_{l=1}^L \mathbf{e}_l \beta_{E_l}}_E + \underbrace{\boldsymbol{\epsilon}}_{\text{noise}}, \tag{3.17}
\end{aligned}$$

where  $\boldsymbol{\beta}'_{G \times E} = [\beta_{G \times E_1}, \beta_{G \times E_2}, \dots, \beta_{G \times E_L}]^T$ ,  $\mathbf{E}$  is the usual  $N \times L$  environment matrix and  $\mathbf{E}\boldsymbol{\beta}'_{G \times E}$  is the aggregate environment.

Comparison of the total genetic component in Eq. 3.17 with the in-sample estimate of the total genetic component (G+G×E) given by Eq. 3.11 (i. e. setting  $\mathbf{x}_\star = \mathbf{x}$  and  $\mathbf{E}_\star = \mathbf{E}$ ), shows that a maximum a posteriori estimate of  $\boldsymbol{\beta}'_{G \times E}$  is given by:

$$\boldsymbol{\beta}'_{G \times E} = \hat{\sigma}_{G \times E}^2 (\mathbf{x} \odot \mathbf{E})^T \hat{\mathbf{K}}^{-1} (\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\alpha}} - \mathbf{x}\hat{\beta}_G) \tag{3.18}$$

and thus the maximum a posteriori estimate of the aggregate environment driving the G×E effect at a variant is given by:

$$\mathbf{E}\boldsymbol{\beta}'_{G \times E} = \hat{\sigma}_{G \times E}^2 \mathbf{E}(\mathbf{x} \odot \mathbf{E})^T \hat{\mathbf{K}}^{-1} (\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\alpha}} - \mathbf{x}\hat{\beta}_G). \tag{3.19}$$

It can be seen that this estimated aggregate environment is equivalent to the G×E component of the estimated allelic effect described by Eq. 3.15.

### 3.2.5 Computational complexities of methods used to explore loci

All of the methods described in this section, used to explore identified loci, are based on fitting the interaction test marginal model described by Eq. 2.27 (see Chapter 2). Noting that this model can be written as:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{W}\boldsymbol{\alpha} + \mathbf{x}\beta_G, [\sigma_{G \times E} \text{diag}(\mathbf{x})\mathbf{E}, \sigma_e \mathbf{E}][\sigma_{G \times E} \text{diag}(\mathbf{x})\mathbf{E}, \sigma_e \mathbf{E}]^T + \sigma_n^2 \mathbf{I}_N), \tag{3.20}$$

the same efficient inference scheme as was used to fit the null model for the StructLMM interaction and joint association tests<sup>79,253</sup> (see Section 2.2.6) can be employed; the rank of the first covariance component is now  $2L$  in place of  $L$ , such that the computational complexity of fitting this model is  $O(4NL^2 + 8L^3)$ .

Whilst the parameter inference of this model still scales linearly in the number of samples, inclusion of the additional random effect component that models  $G \times E$ , results in slower inference than when fitting the null model, described in Section 2.2.5. As a result, it is recommended that these analyses are only performed on loci with significant interaction or joint association effects.

## 3.3 Methods

In this section, I will describe methods that I use throughout this chapter which are not specific to StructLMM. I will start by describing the data pre-processing steps that were applied to the UK Biobank data, followed by details of the calibration assessment and finally, the method used to identify independent and exclusive loci.

### 3.3.1 UK Biobank data processing

The analyses in this chapter have been conducted using the full release of UK Biobank (Application 14069)<sup>215,271</sup>.

#### Phenotype and environment pre-processing

BMI phenotype data is ‘Instance 0’ of the UK Biobank field ‘21001’ (see <http://biobank.ctsu.ox.ac.uk/crystal/> for details). Any individuals with missing BMI data were discarded from the analysis and the remaining BMI values were log transformed<sup>195,338</sup>.

Following Young *et al.*<sup>195</sup>, 20 environmental variables: 9 ordinal dietary variables (‘Oily fish intake’, ‘Non-oily fish intake’, ‘Processed meat intake’, ‘Poultry intake’, ‘Beef intake’, ‘Lamb/mutton intake’, ‘Pork intake’, ‘Cheese intake’ and ‘Salt added to food’), three continuous dietary variables (‘Cooked vegetable intake’, ‘Bread intake’, ‘Tea intake’), three physical activity variables (‘Number of days/week walked 10+ minutes’, ‘Number of days/week of moderate physical activity 10+ minutes’,

‘Number of days/week of vigorous physical activity 10+ minutes’), ‘Alcohol intake frequency’, ‘Sleep duration’, ‘Townsend deprivation index’, ‘Smoking status’ and ‘Time spent watching television’, using the data from ‘Instance 0’ (see <http://biobank.ctsu.ox.ac.uk/crystal/> for details), were selected. Again, following Young *et al.*<sup>195</sup>, for the three continuous dietary variables, I removed values exceeding the 99<sup>th</sup> percentile. For ‘Sleep duration’, I removed the top and bottom percentiles and for each individual calculated the squared deviations from the mean sleep duration, creating an additional environmental variable, ‘Squared sleep duration res.’ (21<sup>st</sup> environmental variable). For ‘Time spent watching television’, less than 0.5 hours of was encoded as 0.5 and I excluded individuals in the upper and lower percentile. I then removed individuals with any missing environmental data.

I further removed any remaining poor quality samples flagged by UK Biobank using the field ‘het.missing.outliers’ (based on heterozygosity and the amount of missing data) and individuals with more than ten 3<sup>rd</sup> degree relatives using the field ‘excess.relatives’ in the released ‘Sample-QC’ file. I then kept only those individuals that were genetically ‘White British’ based on the field ‘in.white.British.ancestry.subset’ and finally removed any remaining individuals that were listed in field ‘ID1’ using the released ‘Relatedness’ file such that there were no relatives (3<sup>rd</sup> degree or closer) included in the analysis. This left a total of 252,188 individuals.

## Generation of principle components

Genetic principle components (PCs) were generated using FlashPCA version 2.0<sup>339</sup>, based on the set of 147,604 SNPs flagged by the field ‘used.in.pca.calculation’ in the released ‘Sample-QC’ file and the 252,188 individuals that passed all QC procedures (described above). Ten PCs were used to control for population structure.

## Imputed genotype QC

Testing was performed using the released imputed genotype data, considering only SNPs that were imputed based on the HRC panel, <http://www.haplotype-reference-consortium.org/site> (there were mapping issues with variants that were imputed using UK10K+1000 Genomes panel array when this analysis was conducted).

A fast bgen reader (<https://github.com/limix/bgen-reader-py>), was used to load in the genotypes and genotype QC procedures were incorporated into the

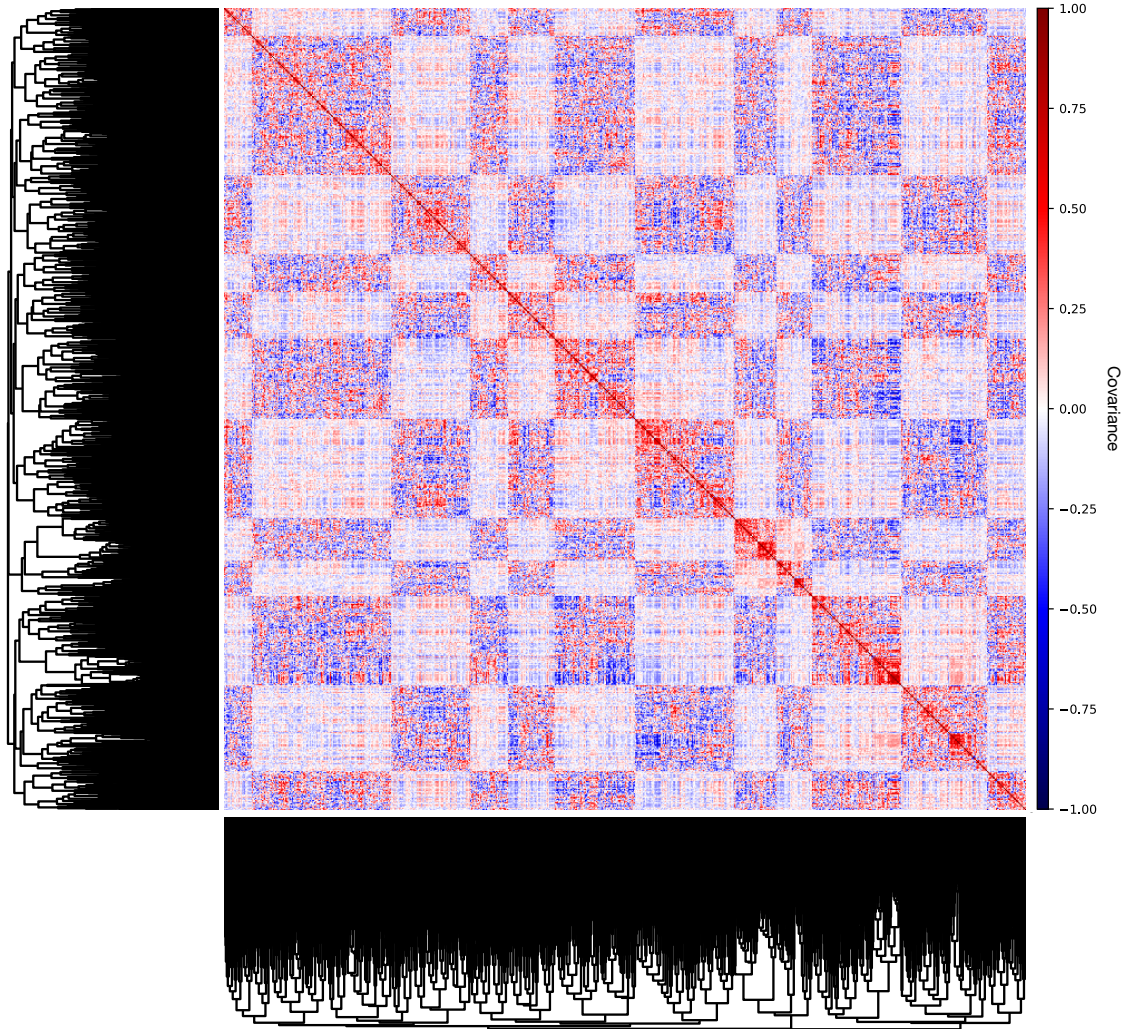
StructLMM interaction and association testing procedures, such that data after intermediate QC steps was not stored. Specifically, I first replaced any genotype-sample probabilities with missing values if the maximum probability across the genotypes (the bgen format provides probabilities of the genotype being homozygous reference, heterozygous or homozygous alternative) was below 0.5. I then calculated the dosages for non-missing sites and removed SNPs if  $> 5\%$  samples had missing dosages,  $MAF < 1\%$ ,  $HWE < 1 \times 10^{-6}$  and INFO score  $r^2 < 0.4$  (using the INFO score provided in the released ‘Imputation MAF+info’ file). 7,515,856 SNPs passed these filters. For all remaining variants that had a maximum genotype-sample dosage probability below 0.5, I calculated a dosage from the provided probabilities as I deemed this to be better than mean imputation. Any remaining dosage-sample pairs with missing data were mean imputed. Genotype dosages were mean centred and standardised.

### Environment covariance, $\Sigma$ , and model covariates

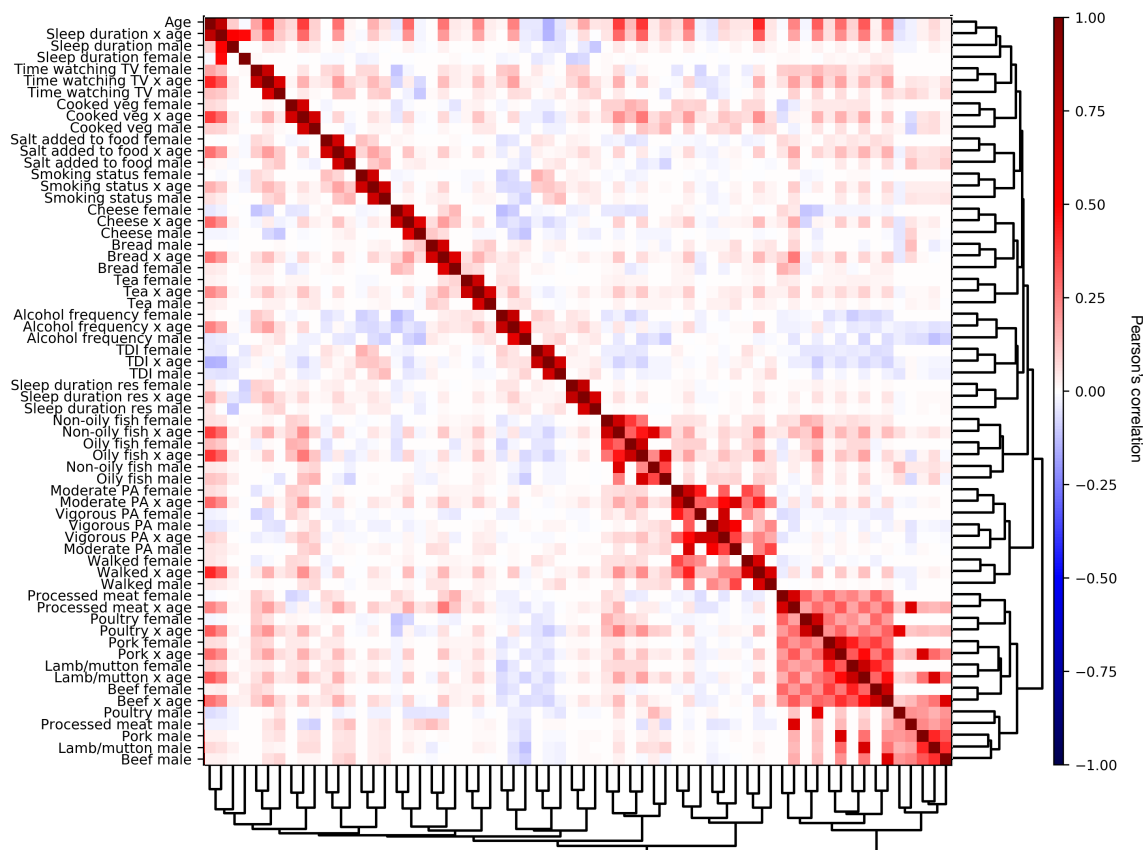
The 21 processed environments (described above) were age and gender adjusted, as described in Section 2.2.4, and age itself added to the environmental matrix,  $\mathbf{E}$ , resulting in a total of 64 environments. Each environmental variable was then mean centred (taking care to mean adjust only environmental values i.e. excluding those set to 0 by gender adjustment) and standardised, followed by a rescaling such that the per-individual variance was 1 (see Section 2.2.4). Fig. 3.1 shows an example of the subsequently generated environment covariance,  $\Sigma$ , based on 5,000 randomly selected individuals and Fig. 3.2 the correlation of the 64 environments across the 252,188 individuals. For downstream analyses (see Section 3.4.4), I did not perform per-individual rescaling to enable direct interpretation (see Section 2.2.4). All models used for testing and downstream analysis include a mean vector, genotype chip (two different chips were used to collect the UK Biobank genotype information), gender,  $age^2$ ,  $age^3$ ,  $gender \times age$ ,  $gender \times age^2$ ,  $gender \times age^3$  and 10 PCs (generated as described above) to account for population structure.

### 3.3.2 Calibration

Calibration was assessed by permuting the 173,297 variants (over the 252,188 individuals) on chromosome 20 such that any true association and interaction signals should be destroyed. The resulting 173,297 P values from interaction or association testing were used to generate QQ plots of the expected negative log P values



**Fig. 3.1 Covariance structure,  $\Sigma$ , of UK Biobank individuals based on 64 environmental variables** | Sample covariance matrix for 5,000 randomly selected individuals, calculated using the 64 environmental variables considered for the analyses presented in this chapter: 12 diet-related factors, three physical activity factors and six lifestyle factors, modelled as gender and age adjusted and age itself. Dark red denotes pairs of individuals with stronger environmental similarity, whilst blue corresponds to negative covariance of environmental similarity (anti correlation). Individuals are ordered using hierarchical clustering.



**Fig. 3.2 Correlation of the 64 environments based on 252,188 UK Biobank individuals** | Shown are correlation coefficients between pairs of environmental variables, considering 12 diet-related factors, three physical activity factors and six lifestyle factors, modelled as gender and age adjusted and age itself. Environments are ordered using hierarchical clustering.

(x-axis) versus the observed negative log P values (y-axis) and inflation parameters,  $\lambda_{GC} = \frac{\log_{10}(m)}{\log_{10}(0.5)}$ , where  $m$  is the median P value of the 173,297 variants tested, were calculated.

### 3.3.3 Defining independent and exclusive loci

To define independent loci from the significantly identified variants (Benjamini-Hochberg FDR<sup>86</sup> adjusted  $P < 0.05$  and  $P < 5 \times 10^{-8}$  for interaction and association tests, respectively), I iteratively (i) selected the most significant variant (using the Benjamini-Hochberg FDR adjusted P values for interaction tests) and (ii) clumped all variants in LD,  $r^2 \geq 0.1$  (calculating LD using the 252,188 UK Biobank individuals that passed sample QC) within  $+/- 500$  kb, until no variant was left. This was done separately for each method considered.

To identify loci that are found exclusively by one of two methods, I (i) for all variants in a clump, identified any variants within  $+/- 500$  kb that were significant for the second method, (ii) calculated LD,  $r^2$  between variants in the selected clump and the significant variants for the second method (if any exist) and (iii) declared the selected clump as exclusively identified if the maximum LD,  $r^2 < 0.1$ .

## 3.4 Results

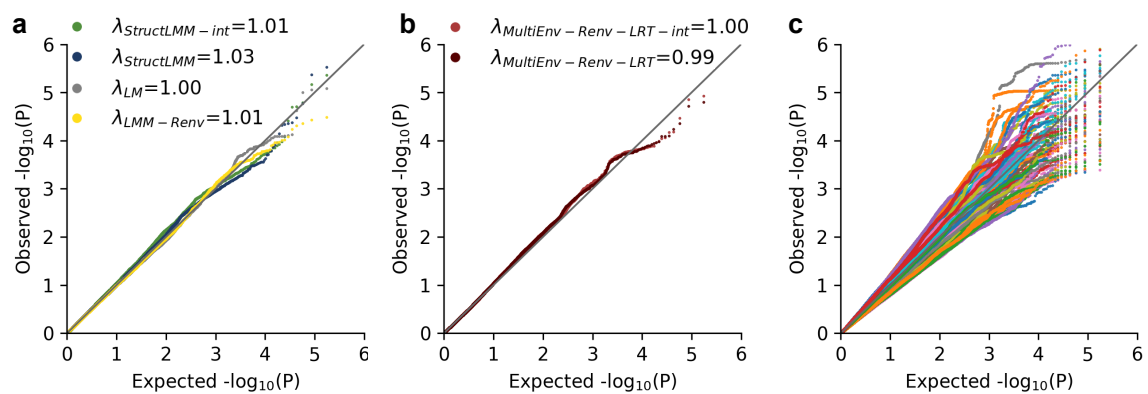
In this section, I first describe the calibration of different interaction and association methods that are used in this chapter. I then outline the key findings when testing for interaction effects between the 64 lifestyle factors and (i) the 97 variants identified by Locke *et al.*<sup>316</sup> and (ii) the variants associated with BMI, identified through a genome-wide scan on the UK Biobank data. This is followed, by the results from a genome-wide joint association test that accounts for possible heterogeneity in variant effects due to  $G \times E$ . Finally, I explore some of the identified interaction loci in further detail using the methods described in Section 3.2.

### 3.4.1 Calibration assessment

To check that the data was pre-processed appropriately, I assessed the empirical calibration of different models used in this chapter (see Section 3.3.2 for details). The StructLMM interaction (StructLMM-int) and association tests, as well as, the



linear models, LMM-Renv and LM (see Section 2.3 for model details) were calibrated (Fig. 3.3a). In addition, the multi-environment interaction (MultiEnv-Renv-LRT-int) and association (MultiEnv-Renv-LRT) methods based on fixed effects (see Section 2.3 for model details) were calibrated in this setting, since the sample size is sufficiently large (Fig. 3.3b; see Section 2.4.4 for a discussion on the settings in which these methods are calibrated). However, I do note that the single environment interaction tests, SingleEnv-Renv-int (see Section 2.3 for model details) show variable levels of statistical calibration (Fig. 3.3c). This has been noted in previous interaction studies<sup>340–344</sup> and may not be symptomatic of uncontrolled confounding factors but instead occur due to minor misspecification of the null model (possibly due to heteroskedasticity in the noise term). These minor misspecifications do not cause a lack of observed calibration for marginal effect scans due to independence of the test statistics across the variants. However, the interaction term is a product of  $\mathbf{e}_l$  and  $\mathbf{x}$ , with the former repeatedly used throughout the empirical calibration experiment, such that the test statistics across the variants are not independent (I refer the reader to Rao *et al.*<sup>344</sup> for further details). Since all other methods, in particular StructLMM-int and MultiEnv-LRT-int that fit identical null models, were calibrated, it is unlikely that the observed inflation and deflation is due to uncorrected confounding. I did not attempt to correct this mis-calibration, as the SingleEnv-Renv-int tests are not the focus of the work presented in this chapter and adjustment is not straight forward or without further potential problems<sup>344</sup>. These results again highlight that StructLMM is more robust than other interaction methods.



**Fig. 3.3 Calibration of interaction and association tests for BMI on UK Biobank data** | QQ plots of negative log P values from different interaction and association tests applied to UK Biobank BMI phenotype data based on permuted genetic variants (chromosome 20). (a) QQ plots for StructLMM-int (green), StructLMM (blue), LM (grey) and LMM-Renv (yellow). (b) QQ plots for MultiEnv-Renv-LRT-int (salmon) and MultiEnv-Renv-LRT (red). (c) QQ plots for SingleEnv-Renv-int, for each of 64 considered environmental variables.

### 3.4.2 Interaction test results

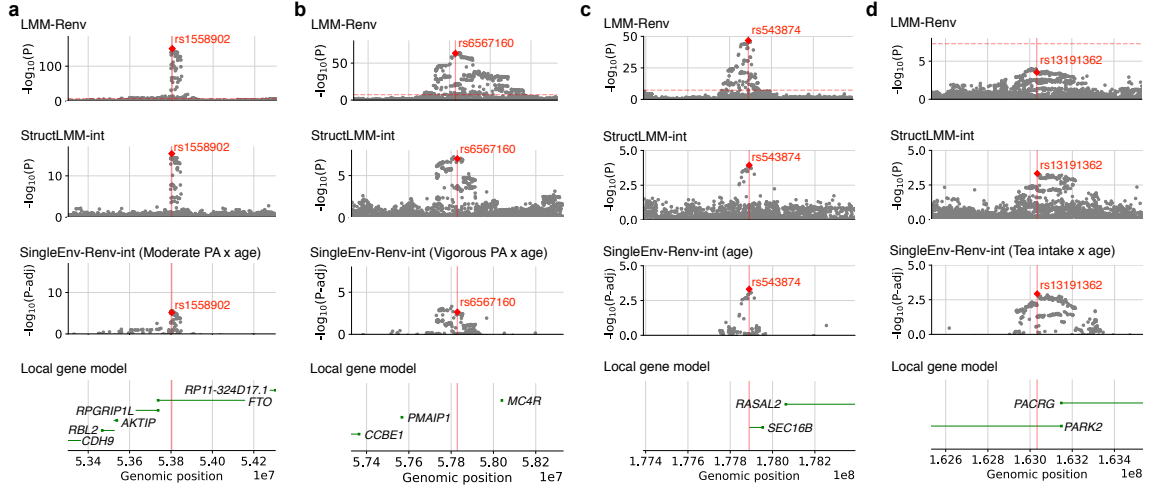
#### Interaction test results at the 97 GIANT variants

The majority of previous interaction studies for BMI, have focussed on variants (or subsets of variants) that were significantly associated with BMI in marginal association studies. At the time of conducting these analyses, the largest study for which associations with BMI had been tested, was the meta-analysis performed by the GIANT consortium<sup>316</sup> that identified 97 loci associated with BMI ( $P < 5 \times 10^{-8}$ ). Thus, I initially tested for  $G \times E$  with the 64 lifestyle based environments (as described in Section 3.3.1) at these 97 loci (all of which passed the imputed genotype QC thresholds described in Section 3.3.1), using the 252,188 individuals that passed sample QC (see Section 3.3.1).

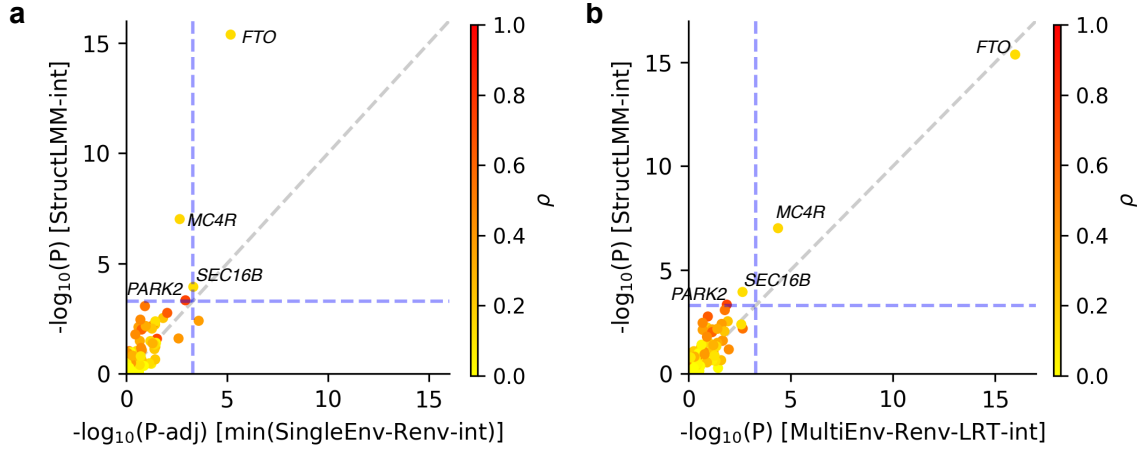
StructLMM-int identified 4 significant interactions at a FWER 5%. One of these was at the *FTO* locus (*rs1558902*,  $P = 4.23 \times 10^{-16}$ ,  $\rho = 0.138$ , Fig. 3.4a), previously reported to interact with physical activity<sup>195–199,206,213,324</sup>, diet<sup>195,196,200–205,214</sup>, alcohol consumption<sup>195</sup>, squared residuals of sleep duration<sup>195</sup>, and smoking for men aged 40 – 60<sup>196</sup>. A second, was at the *MC4R* locus (*rs6567160*,  $P = 9.78 \times 10^{-8}$ ,  $\rho = 0.156$ , Fig. 3.4b), for which one previous study has reported an interaction effect with physical activity for women aged 20 – 40<sup>196</sup>. Another, was at the *SEC16B* locus (*rs543874*,  $P = 1.15 \times 10^{-4}$ ,  $\rho = 0.133$ , Fig. 3.4c), for which there is suggestive evidence of an interaction with physical activity in Europeans<sup>213</sup> and Hispanics<sup>324</sup> and also diet<sup>214</sup> ( $P = 0.025$ ,  $P = 0.003$  and  $P = 0.02$  prior to multiple testing corrections for 12, 37 and 32 variants, respectively). The final interaction was at the *PARK2* locus (*rs13191362*,  $P = 4.69 \times 10^{-4}$ ,  $\rho = 0.736$ , Fig. 3.4d), which has been linked to BMI change in a ten year follow up study<sup>331</sup>.

For comparison, only two of these four loci, *FTO* and *SEC16B* were significant when testing for interactions with a single environment, using a Bonferroni correction to account for testing 64 environments per variant (P-adj; SingleEnv-Renv-int, see Section 2.3 for model details). In addition, these interactions were more significant with StructLMM-int ( $P = 4.23 \times 10^{-16}$  and  $P = 1.15 \times 10^{-4}$  at *FTO* and *SEC16B*, respectively) than with SingleEnv-Renv-int (P-adj =  $6.76 \times 10^{-6}$  and P-adj =  $4.80 \times 10^{-4}$  at *FTO* and *SEC16B*, respectively; Fig. 3.5a, Fig. 3.4a, b). Similarly, when testing for interactions using a multi-environment model based on fixed effects (MultiEnv-Renv-LRT-int, see Section 2.3 for model details), only two of these four loci, *FTO* and *MC4R* were identified (Fig. 3.5b).

I also examined the results when using the more relaxed 5% FDR (Benjamini-Hochberg



**Fig. 3.4 Local Manhattan plots for the four interaction loci identified by StructLMM-int** | Local Manhattan plots of interactions identified by StructLMM-int (5% FWER) at (a) *FTO*, (b) *MC4R*, (c) *SEC16B* and (d) *PARK2* respectively. From top to bottom: LMM-Renv association test, StructLMM interaction test, SingleEnv-Renv interaction test for the environment with the most significant  $G \times E$  effect at the respective GIANT variant (P values Bonferroni adjusted to account for the number of tested environments, P-adj) and local gene models. The red vertical line and diamond symbol indicates the position of the GIANT variant as annotated by Locke *et al.*<sup>316</sup>.



**Fig. 3.5 Comparison of interaction results at the 97 GIANT loci** | Scatter plot of negative log P values from  $G \times E$  interaction tests at 97 GIANT variants, considering (a) single environment fixed effect  $G \times E$  tests, plotting the result for the environment with the most significant  $G \times E$  effect at a given variant (SingleEnv-Renv-int, x-axis; P values Bonferroni adjusted for the number of tested environments, P-adj) and (b) multi-environment fixed effect  $G \times E$  tests (MultiEnv-Renv-LRT-int, x-axis) versus the StructLMM interaction test (StructLMM-int, y-axis). Dashed lines correspond to  $\alpha < 0.05$ , Bonferroni adjusted for the number of tested variants. Colour denotes the estimated fraction of the total genetic variance due to  $G \times E$  ( $\rho$ , see Section 3.2.1 for details), where yellow/red corresponds to variants with low/high  $G \times E$  components.

adjustment<sup>86</sup>) threshold. This results in a greater difference in the number of loci identified by each method, where StructLMM-int identifies an additional seven loci (11 in total), *LMX1B*, *TOMM40*, *MTCH2*, *TLR4*, *UBE2E3*, *ADCY3* and *MIR548X2*, to the four already mentioned, whilst SingleEnv-Renv-int identifies an additional three loci (including *MC4R* and *PARK2*; six in total) and MultiEnv-Renv-LRT-int identifies no additional loci (two in total).

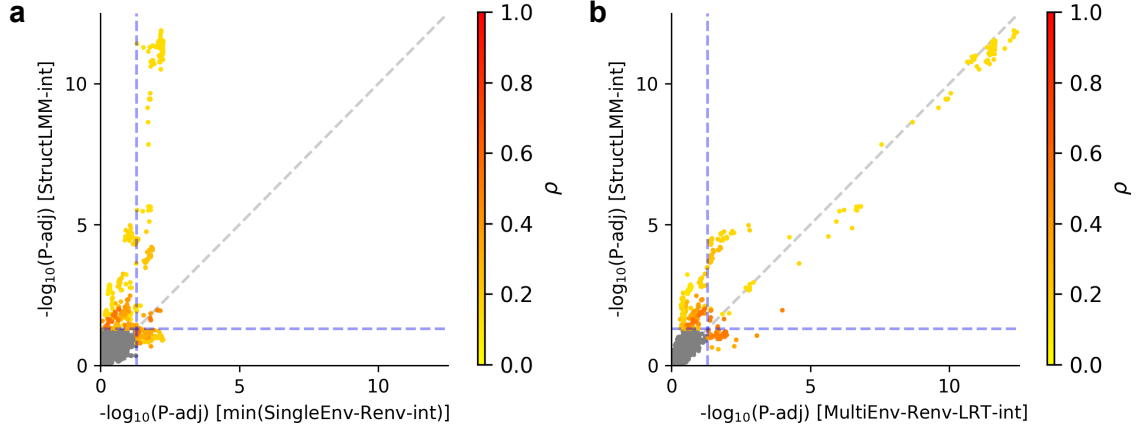
I refer the reader to Supplementary Table 3 in Moore *et al.*<sup>245</sup> for a tabular version of results for all GIANT variants.

## Interaction test results based on marginal association results in UK Biobank

As an alternative filtering strategy, I performed a genome-wide association scan, using LMM-Renv (see Section 2.3 for model details) with the set of 252,188 UK Biobank individuals and 7,515,856 SNPs that pass sample and genotype QC (see Section 3.3.1), selecting 17,606 significantly associated variants ( $P < 5 \times 10^{-8}$ ). I subsequently tested for  $G \times E$  at these variants with StructLMM-int, SingleEnv-Renv-int and MultiEnv-Renv-LRT-int<sup>†</sup>, using a 5% FDR (Benjamini-Hochberg adjustment<sup>86</sup>) to identify significantly interacting variants. StructLMM-int, SingleEnv-Renv-int and MultiEnv-Renv-LRT-int identified 451, 309 and 273 significant variants (Fig. 3.6), corresponding to 23, 11 and 9 significant loci, respectively ( $+/- 500$  kb, LD  $r^2 < 0.1$ , see Section 3.3.3). As well as identifying previously associated GIANT loci, including *FTO*, *MC4R* and *SEC16B*, StructLMM-int also identified interactions at nine loci that were  $> 500$  kb from the 97 reported GIANT loci<sup>316</sup> (see Supplementary Table 3 in Moore *et al.*<sup>245</sup> for a tabular version of results). This includes a locus (lead variant *rs13264668*), only identified by StructLMM-int ( $P = 1.10 \times 10^{-3}$ ,  $\rho = 0.443$ ), that is upstream of the *MSRA* gene and is  $\approx 25$  kb from a variant associated with waist circumference<sup>345,346</sup>, early onset obesity<sup>347</sup> and in addition was found to be significantly associated with BMI trajectories in female children<sup>348</sup>.

---

<sup>†</sup>noting that the marginal association and interaction tests are independent, see<sup>229</sup> for full details



**Fig. 3.6 Comparison of interaction results at genome-wide significant variants in UK Biobank** | Scatter plot of negative log Benjamini-Hochberg adjusted P values (P-adj) from  $G \times E$  interaction tests at 17,606 variants that were significantly associated with BMI using LMM-Renv ( $P < 5 \times 10^{-8}$ ) in UK Biobank, considering (a) single environment fixed effect  $G \times E$  tests, plotting the result for the environment with the most significant  $G \times E$  effect at a given variant (SingleEnv-Renv-int, x-axis; P values Bonferroni adjusted for the number of tested environments) and (b) multi-environment fixed effect  $G \times E$  tests (MultiEnv-Renv-LRT-int, x-axis) versus the StructLMM interaction test (StructLMM-int, y-axis). Dashed lines correspond to  $P\text{-adj} < 0.05$  and variants that exceed this threshold are coloured by the estimated fraction of the total genetic variance due to  $G \times E$  ( $\rho$ , see Section 3.2.1 for details), where yellow/red corresponds to variants with low/high  $G \times E$  components.

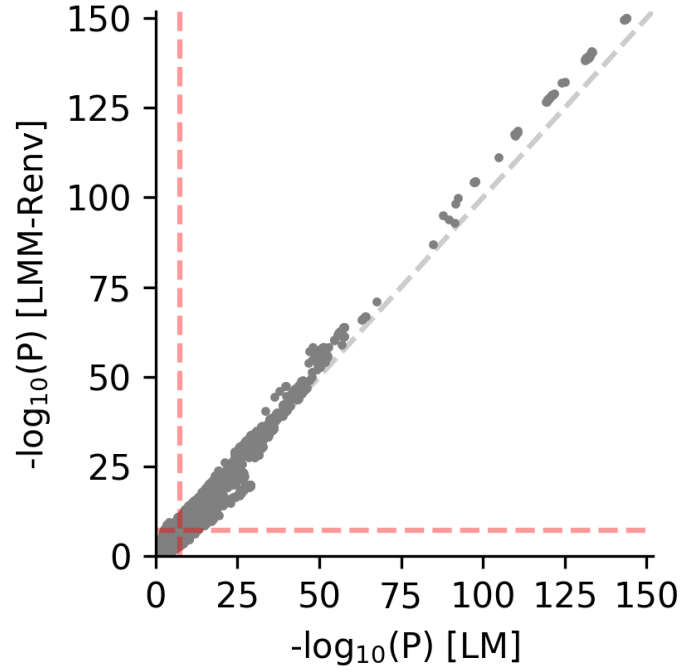
### 3.4.3 Genome-wide association results

As already outlined in Chapter 2, the StructLMM framework can also be used to perform a joint association test, i.e. test for associations whilst accounting for possible heterogeneity in the variant effect across individuals due to  $G \times E$ . Using the same set of 64 lifestyle factors and 252,188 individuals as for the interaction analyses, I tested for associations at the 7,515,856 low frequency and common SNPs that pass QC ( $MAF > 1\%$ ; see Section 3.3.1 for data pre-processing details).

For comparison, I also performed a genome-wide association scan using linear models, LMM-Renv and LM. The former accounts for additive environment effects using a random effect term such that the null model is identical to that of StructLMM (see Section 2.3 for model details) facilitating direct assessment of the impact of modelling  $G \times E$  effects.

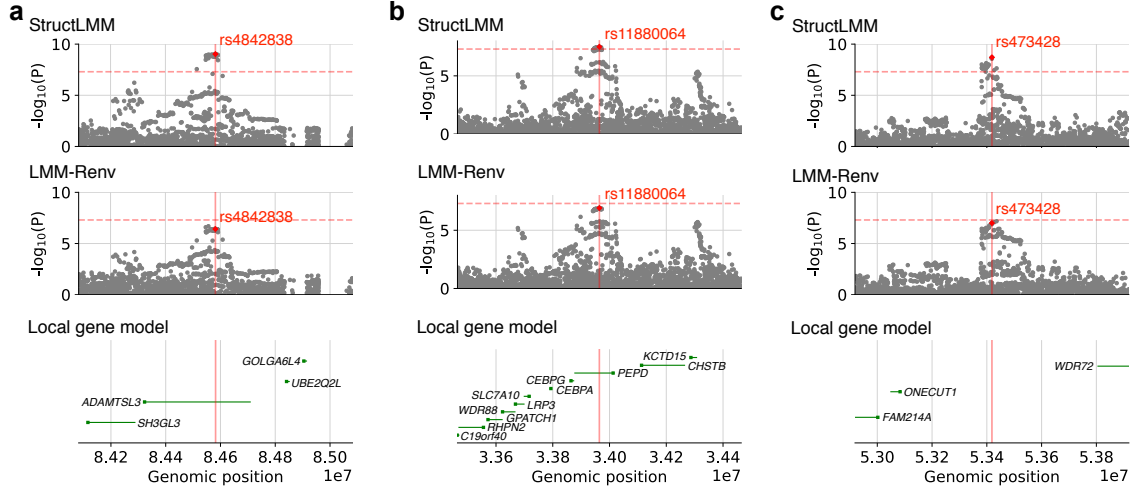
I first note that the linear models LMM-Renv and LM identify slightly different sets of associated loci ( $+/- 500$  kb,  $LD r^2 < 0.1$ , see Section 3.3.3; Fig. 3.7). LMM-Renv, identified 327 loci of which 14.37% were not detected by the LM and the LM, identified 379 loci, of which 25.59% were not detected by LMM-Renv (see Supplementary Table 4 in Moore *et al.*<sup>245</sup> for a tabular version of all significantly associated loci). These results agree with those seen in the simulation experiments (Section 2.4.5), where LMM-Renv was better powered than the LM when environments were correlated with the phenotypic outcome (Fig. 2.4) but the LM can outperform LMM-Renv when the environments were heritable (Fig. 2.10).

Despite this, there were 23 loci found exclusively by StructLMM (351 loci in total,  $+/- 500$  kb,  $LD r^2 < 0.1$ , see Section 3.3.3; see Supplementary Table 4 in Moore *et al.*<sup>245</sup> for a tabular version of all significantly associated loci). One such locus lies in the *ADAMTSL3* gene (lead variant *rs4842838*,  $P$  StructLMM =  $9.35 \times 10^{-10}$ ,  $P$  LMM =  $3.83 \times 10^{-7}$ ,  $P$  LM =  $2.37 \times 10^{-5}$ ,  $\rho = 0.576$ ; Fig. 3.8a) and codes for a glycoprotein<sup>349</sup>. Other variants within this gene have been linked to BMI-related traits, including lean body mass<sup>350</sup>, WC<sup>351</sup> and hip circumference adjusted for BMI<sup>352</sup>. A second interesting locus, lies in the *PEPD* gene (lead variant *rs11880064*,  $P$  StructLMM =  $5.23 \times 10^{-9}$ ,  $P$  LMM =  $1.22 \times 10^{-7}$ ,  $P$  LM =  $8.46 \times 10^{-8}$ ,  $\rho = 0.525$ ; Fig. 3.8b) and codes for a protein involved in the final stage of degradation of endogenous and dietary proteins. Several additional *PEPD* genetic variants have been associated with adiponectin<sup>353,354</sup>, fasting insulin adjusted for BMI<sup>355</sup>, HDL cholesterol<sup>356</sup>, triglycerides<sup>356,357</sup>, type 2 diabetes<sup>358</sup>, WC adjusted for body mass<sup>352</sup> and WHR<sup>352</sup>. A final example, lies upstream of *ONECUT1* and downstream of *WDR72* (lead variant *rs473428*,  $P$  StructLMM =  $1.95 \times 10^{-9}$ ,  $P$  LMM =  $1.01 \times$



**Fig. 3.7 Comparison of LMM-Renv and LM results on UK Biobank data**  
 | Scatter plot of genome-wide negative log P values from the LM association test, without accounting for additive environmental effects (x-axis), versus an LMM association test that accounts for additive environmental effects using the same random effect component as used in StructLMM (LMM-Renv, y-axis). Dashed lines indicate genome-wide significance at  $P < 5 \times 10^{-8}$ .

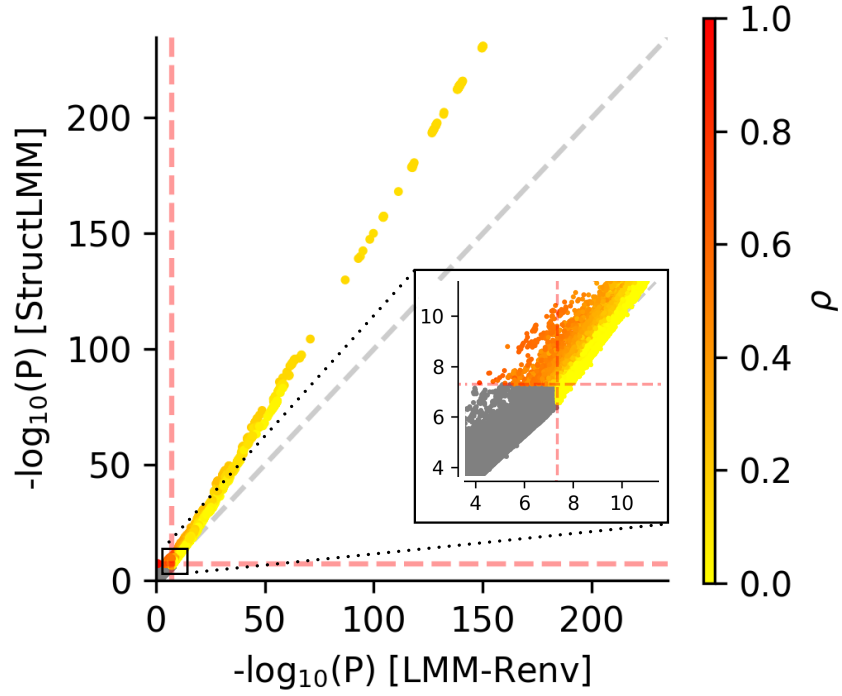
$10^{-7}$ ,  $P_{LM} = 1.26 \times 10^{-7}$ ,  $\rho = 0.415$ ; Fig. 3.8c). *ONECUT1* stimulates the production of liver expressed genes and can inhibit glucocorticoid-stimulated gene transcription<sup>358</sup> and genetic associations with BMI<sup>359</sup>, HDL cholesterol<sup>357</sup>, lipids<sup>357</sup> and triglycerides<sup>357</sup> were reported in early GWAS but have not reached genome-wide significance in more recent studies. This could be due to  $G \times E$  effects varying across different aggregated cohorts or due to differences in trait transformation.



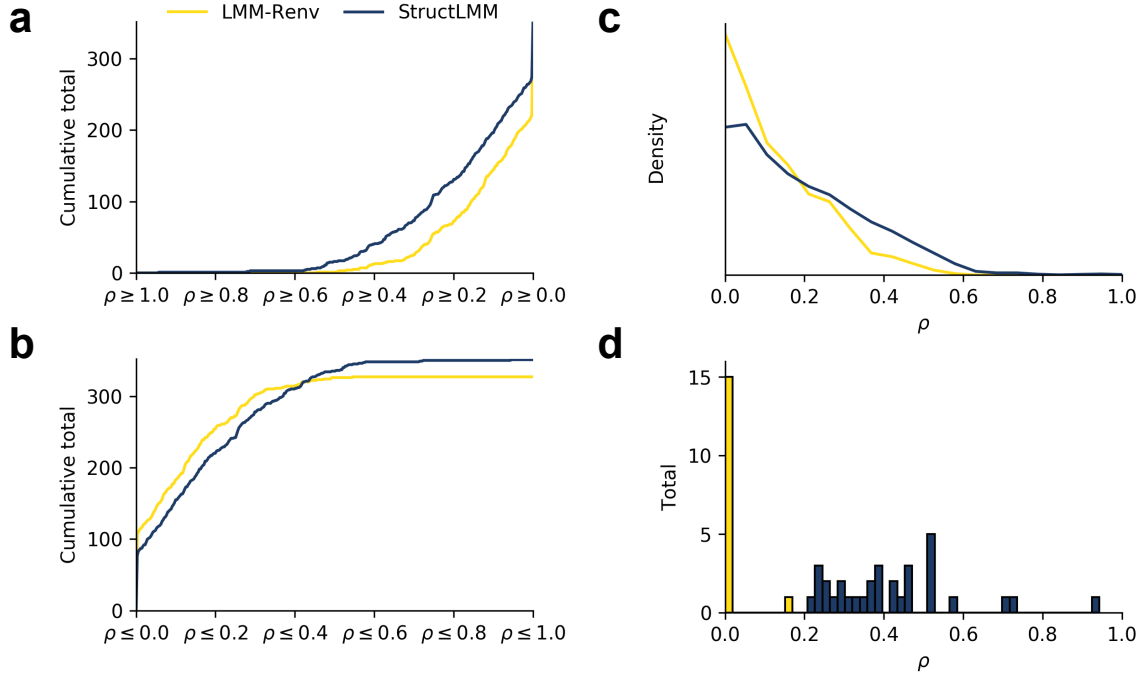
**Fig. 3.8 Local Manhattan plots for three association loci exclusively identified by StructLMM in UK Biobank** | Shown are three associations exclusively identified by the StructLMM association test, with the red vertical line indicating the position of the StructLMM lead variant. From top to bottom: Manhattan plot of negative log P values from StructLMM, LMM-Renv and the local gene models, for (a) *ADAMTSL3*, (b) *PEPD* and (c) *ONECUT1*.

Comparison of  $\rho$  (see Section 3.2.1 for details) at variants (Fig. 3.9) and loci ( $\pm 500$  kb,  $LD\ r^2 < 0.1$ ; see Section 3.3.3 for details) identified by StructLMM (351 in total, 32 not identified by LMM-Renv) and LMM-Renv (327 in total, 16 not identified by StructLMM; Fig. 3.10), reveals that as expected (and in agreement with simulation experiments, Section 2.4, Fig. 2.4a), StructLMM tends to identify loci with a greater extent of  $G \times E$  (larger values of  $\rho$ ), whereas LMM-Renv specific loci tend to have no or little evidence for effect size heterogeneity due to  $G \times E$ . As already mentioned in Chapter 2, the latter can be explained by the fact that StructLMM will be slightly less powered than LMM-Renv when no or very little  $G \times E$  is present since StructLMM is penalised for testing multiple values  $\rho$ . These results indicate that the StructLMM joint association test can be used to identify additional loci, with a  $G \times E$  component.



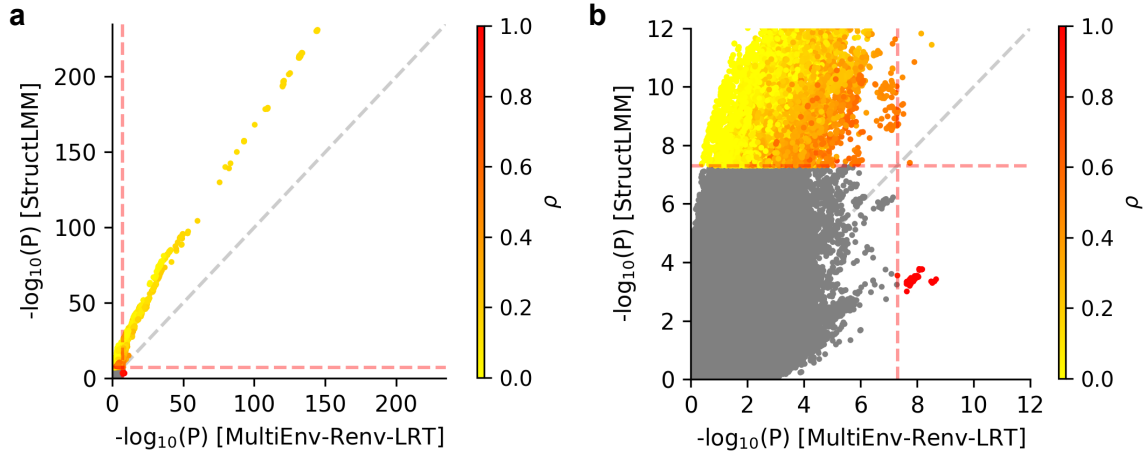


**Fig. 3.9 Comparison of StructLMM and LMM-Renv association tests on UK Biobank data at all 7,515,856 tested variants** | Scatter plot of genome-wide negative log P values from LMM association test (LMM-Renv, x-axis) versus the StructLMM association test (y-axis). Dashed lines indicate genome-wide significance at  $P < 5 \times 10^{-8}$  and colour denotes the estimated extent of heterogeneity ( $\rho$ ), where yellow/red corresponds to variants with low/high  $G \times E$  components. The inset displays a zoom-in view of variants close to genome-wide significance.



**Fig. 3.10** Distribution of the estimated extent of  $G \times E$  for significant loci identified by the StructLMM association test and LMM-Renv on UK Biobank data | Cumulative number of significant associations ( $P < 5 \times 10^{-8}$ ,  $\pm 500$  kb, LD  $r^2 < 0.1$ ) identified by LMM-Renv (yellow,  $N=327$  loci) and StructLMM (blue,  $N=351$  loci) in decreasing (a) and increasing (b) order of the estimated extent of  $G \times E$  ( $\rho$ ). (c) Distribution of the fraction of genetic variance due to  $G \times E$  ( $\rho$ ) for loci identified by LMM-Renv (yellow) and StructLMM (blue). (d) Histogram of the fraction of genetic variance due to  $G \times E$  ( $\rho$ ), considering the subset of loci exclusively identified by either approach: LMM-Renv (yellow, total 16), StructLMM (blue, total 32).

Finally, I compared StructLMM to the multi-environment fixed effect based test, MultiEnv-Renv-LRT (see Section 2.3 for details), which demonstrated that StructLMM was able to identify a much greater number of significantly associated variants (17,630 versus 2,037, Fig. 3.11). The difference in association  $P$  values between StructLMM and MultiEnv-Renv-LRT (Fig. 3.11) is related to the estimated value of  $\rho$ . Again, these results are in agreement, with the simulation experiments (Section 2.4), where power differences between StructLMM and MultiEnv-Renv-LRT become smaller, as the simulated value of  $\rho$  increases (see Fig. 2.6).



**Fig. 3.11 Comparison of multi-environment association tests on UK Biobank data** | (a) Scatter plot of negative log P values from the MultiEnv-Renv-LRT association test (x-axis) versus the StructLMM association test (y-axis). Dashed lines indicate genome-wide significance at  $P < 5 \times 10^{-8}$  and colour denotes the estimated fraction of genetic variance due to  $G \times E$  ( $\rho$ , see Section 3.2.1 for details), with (b) displaying a zoom-in view of variants close to genome-wide significance.

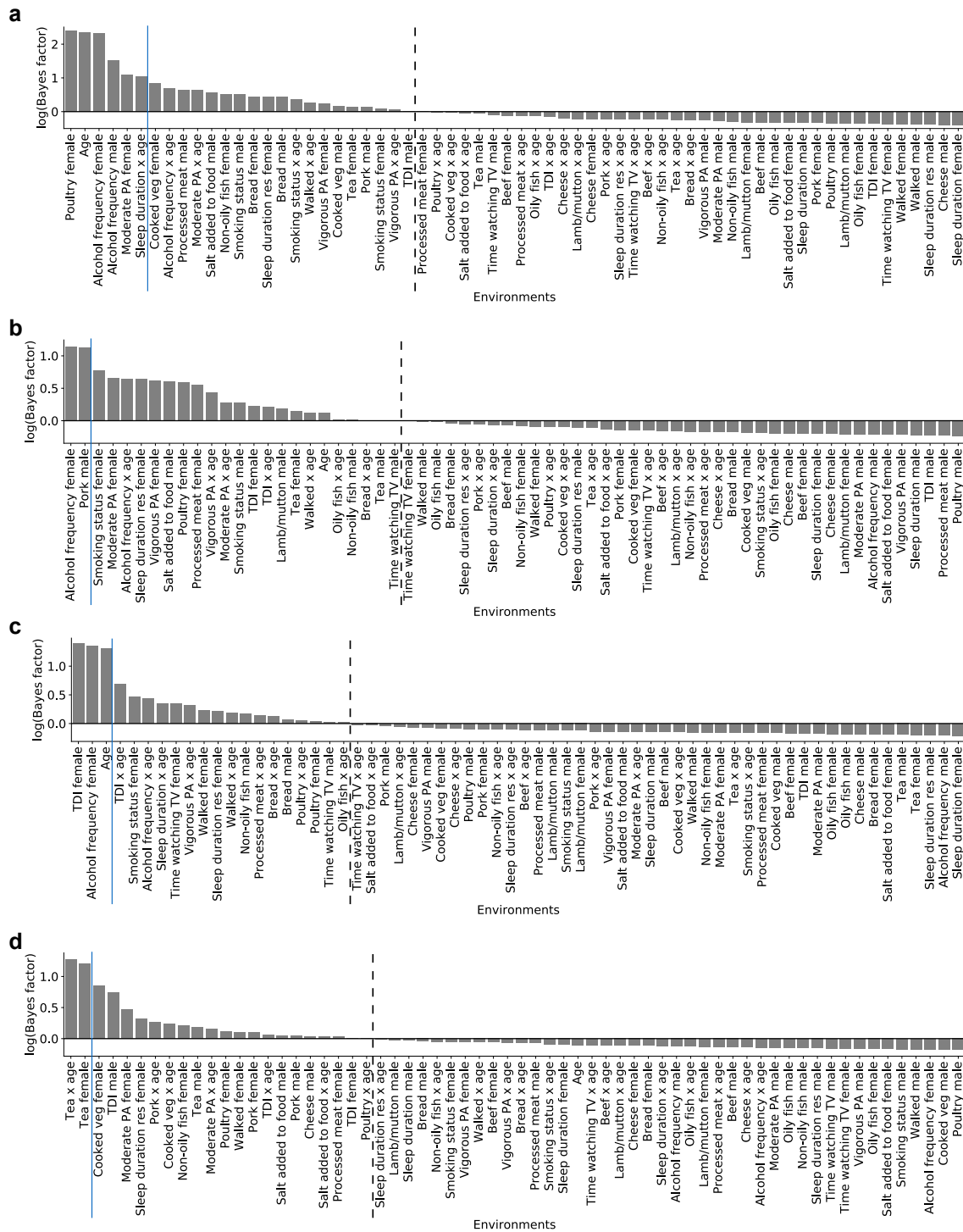
### 3.4.4 Exploration of significant loci

As well as the identification of variants, one of the principle objectives for conducting  $G \times E$  analyses, is to improve the understanding of functional mechanisms that result in increased (or decreased) trait and disease risk.

#### Exploration of environments driving observed $G \times E$ effects

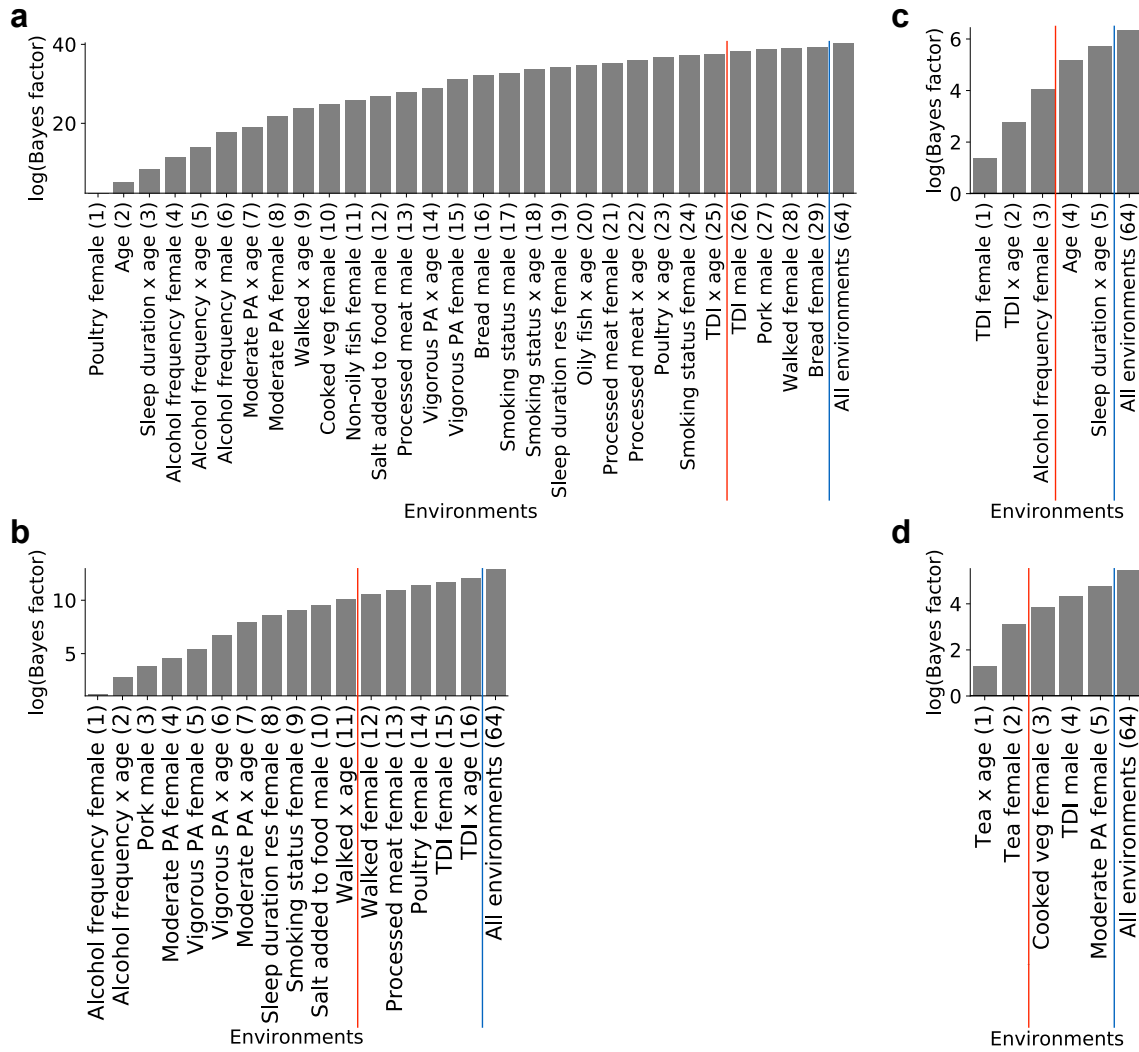
I first explore the environments driving the identified  $G \times E$  effect at the four GIANT loci (FWER 5%) using Bayes factors (see Section 3.2.2 for details). Initially, I examine the evidence for each environment contributing to the observed  $G \times E$  effect, noting that at these four loci no environment displays strong evidence (Bayes factor  $> 3$  based on the log Kass Rafferty scale<sup>333</sup>) of driving the observed interaction effect. However, multiple environments show positive (Bayes factor  $> 1$ , based on the log Kass Rafferty scale<sup>333</sup>) and putative evidence (Bayes factor  $> 0$ ), with the number of such putative environments ranging from 20 to 25, indicative of multi environment  $G \times E$  effects (Fig. 3.12).

Since the environments are not independent of one another, I used a greedy backward elimination procedure (see Section 3.2.2 for details) to identify a potential set of the driving environments at the four loci. This analysis again revealed that multiple



**Fig. 3.12 Evidence for each environment contributing to the identified  $G \times E$  interaction effects using Bayes factors** | Relevance of individual environmental variable for  $G \times E$  effects at four loci, showing Bayes factors between models containing all 64 environments and models with a single environmental variable removed. Shown are results for (a) *FTO*, (b) *MC4R*, (c) *SEC16B* and (d) *PARK2*, ordered by Bayes factor, with environments for which there was evidence of contribution to the  $G \times E$  effect left of the black dashed lines whilst those that showed no evidence of contribution to the right and environments with positive evidence based on the log Kass Rafferty scale<sup>333</sup> left of the blue lines.

environments contribute to the observed interaction effect with the number of environments required to explain the  $G \times E$  effects (strong evidence based on the log Kass Rafferty scale<sup>333</sup>, Section 3.2.2) ranging from 2 at *PARK2* to 25 at *FTO* (Fig. 3.13). The environments selected also varied across the loci, with *PARK2* interacting with ‘tea intake’, an environment that is not selected at any other loci. Interestingly, at *MC4R*, I identified gender specific (women only) and age dependent interaction effects with physical activity (Fig. 3.13b), in agreement with the only previous study that found an interaction effect at this locus<sup>196</sup>. However, this and the aforementioned analysis that removed individual environments, suggests that other environmental factors may be of greater importance, thus underlining the benefits of multi-environment interaction tests.



**Fig. 3.13 Exploration of the environments contributing to the identified  $G \times E$  interaction effects using Bayes factors** | Cumulative evidence of environmental variables contributing to  $G \times E$  at (a) *FTO*, (b) *MC4R*, (c) *SEC16B* and (d) *PARK2* showing Bayes factors between the full model and models with increasing numbers of environmental variables removed using (greedy) backward elimination. For comparison, shown is the total evidence of all environmental variables. The additional environment that is removed at each elimination step is labelled on the x-axis with the total number of environmental variables removed in the considered models shown in brackets. I stopped selecting environments when there was positive evidence based on the log Kass Rafferty scale<sup>333</sup> that I had selected a full set of environments that can drive the observed  $G \times E$  effect and this set of environments are those displayed left of the blue line. Sets of environments required for strong evidence, again based on the log Kass Rafferty scale<sup>333</sup> are shown left of the red line.

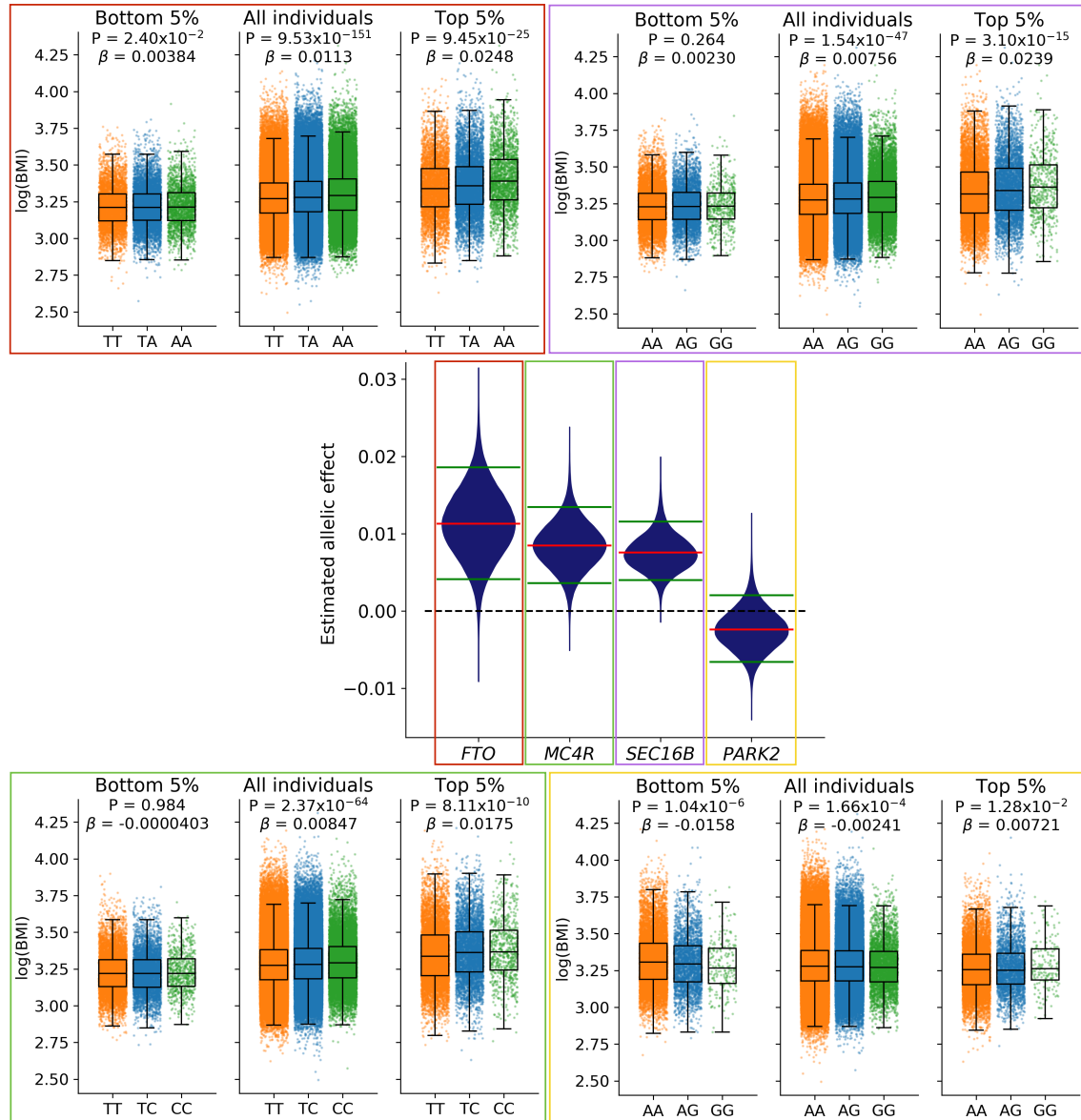
## Identification of individuals at increased and decreased disease risk

Instead of identifying specific environments that drive the observed  $G \times E$  effects, it is perhaps more pertinent (due to the correlation between environments) to explore the aggregate effect of all environments, thereby identifying environmental states (and equivalently individuals if they carry the risk increasing allele) that are potentially at the extremes of the trait or disease risk spectrum. To demonstrate this, I first estimated in-sample per-individual allelic effects at the four GIANT loci with significant  $G \times E$  effects (FWER 5%, Fig. 3.14; see Section 3.2.3 for details). I then selected the 5% of individuals (12,610 individuals) at the extremes of the estimated spectrum<sup>†</sup> and subsequently obtained estimates of the variant effects in these extreme strata using LMM-Renv (Fig. 3.14). For comparison, I also estimated the population variant effect (using all 252,188 UK Biobank individuals), again using LMM-Renv (Fig. 3.14). These results indicate that whilst for *FTO*, *MC4R* and *SEC16B* the aggregate environment only acts to exacerbate the genotype effect, at *PARK2* there is suggestive evidence that different environmental profiles can interact with the genotype with opposite directions of effect on BMI.

To assess the validity of identifying individuals within the population at increased and decreased trait risk, using the per-individual allelic effect in-sample estimates, I performed a hold-out validation experiment. For this, I used the 11 GIANT loci with significant  $G \times E$  effects (FDR 5%, Benjamini-Hochberg adjusted<sup>86</sup>), randomly splitting the population into two groups of equal size (126,094 individuals), a training and a testing set. I estimated per-individual allelic effects for the training set in an identical manner to those generated for the full set of individuals (see Fig. 3.14), again identifying the 5% of individuals (6,305 individuals) with the highest and lowest estimated allelic effects and tested for associations in these extreme strata using LMM-Renv, thus identifying strata that were nominally significant ( $P < 0.05$ ; Fig. 3.15a). I then made out-of-sample predictions for the test set of individuals (which required using only their environmental information, see Section 3.2.3 for details; Fig. 3.15b). To assess the validity of the predictions, I compared the predicted mean allelic effects (yellow crosses, Fig. 3.15b; x-axis, Fig. 3.15c) with the in-sample estimates of the allelic effects obtained using LMM-Renv (at this stage using the genotype and phenotype information of the test individuals; y-axis, Fig. 3.15c), for the nominally significant strata identified based on the training set of individuals. To ensure that the persistent genetic effect was not dominating these

---

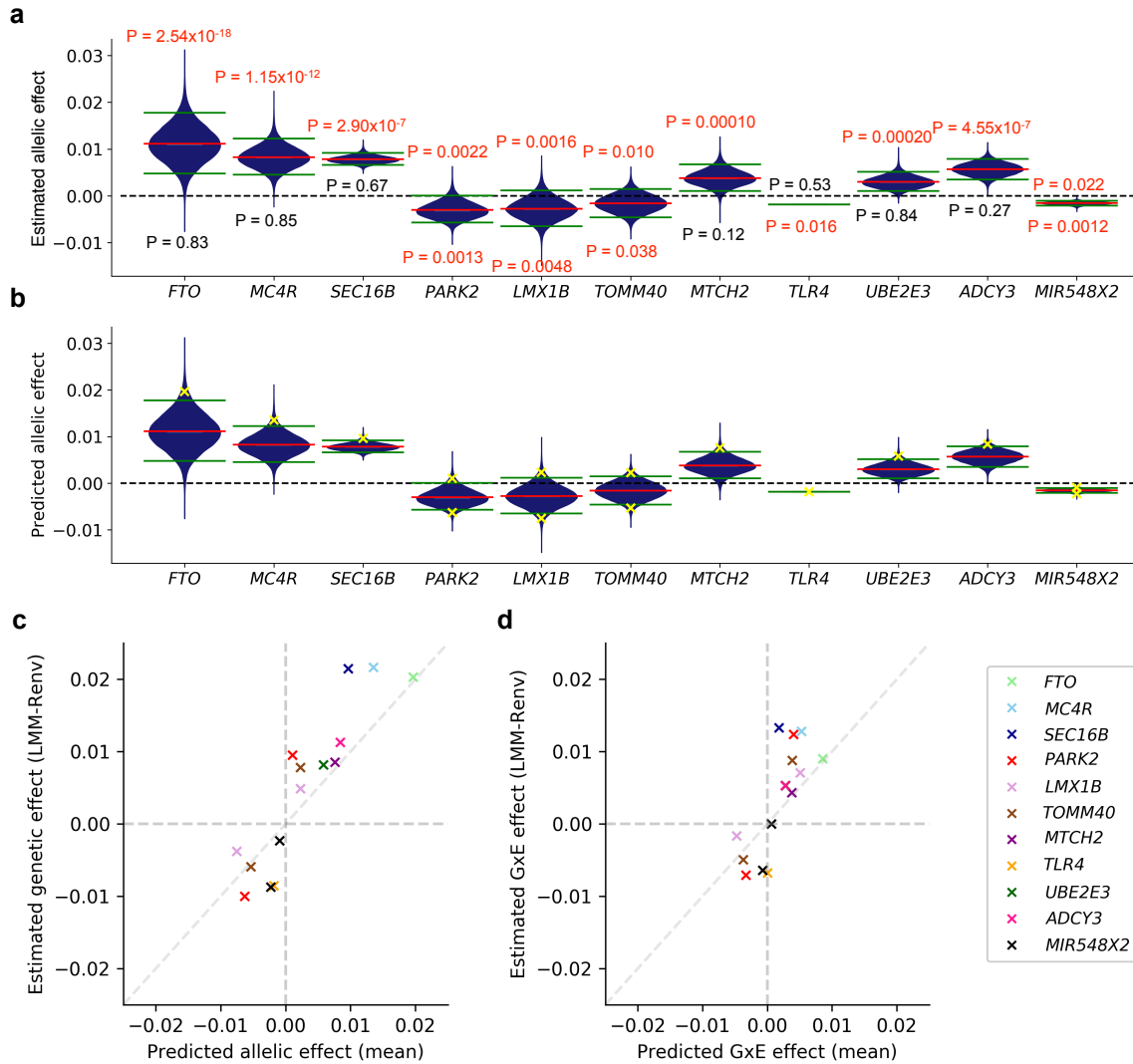
<sup>†</sup>noting that these per-individual allelic effects are the sum of the estimated persistent effect and the estimated aggregate environment effect, this is equivalent to selecting individuals with the most extreme estimated aggregate environment effects



**Fig. 3.14 Exploration of the aggregate environment effect at identified  $G \times E$  loci** | Violin plot (centre) showing the distributions of the in-sample allelic effect size estimates (effect of heterozygous versus homozygous reference carriers for the environmental states observed in the population) on BMI for the four GIANT variants with  $G \times E$  (FWER 5%). Estimated persistent genetic effects are shown by the red bars and the green bars indicate the top and bottom 5% quantiles of variation in effect sizes due to  $G \times E$ . Clockwise from top left: box plots of the genotype (x-axis) versus the phenotype (y-axis) for *FTO*, *SEC16B*, *PARK2* and *MC4R*, for individuals at the extremes of the risk spectrum (individuals in the top and bottom 5% quantiles based on the estimated aggregate environments, or equivalently those individuals that lie above and below the green bars in the violin plot) and equivalently for all individuals in the population. The displayed P values and mean estimated allelic effects ( $\beta$ ) for different population strata are calculated using LMM-Renv.

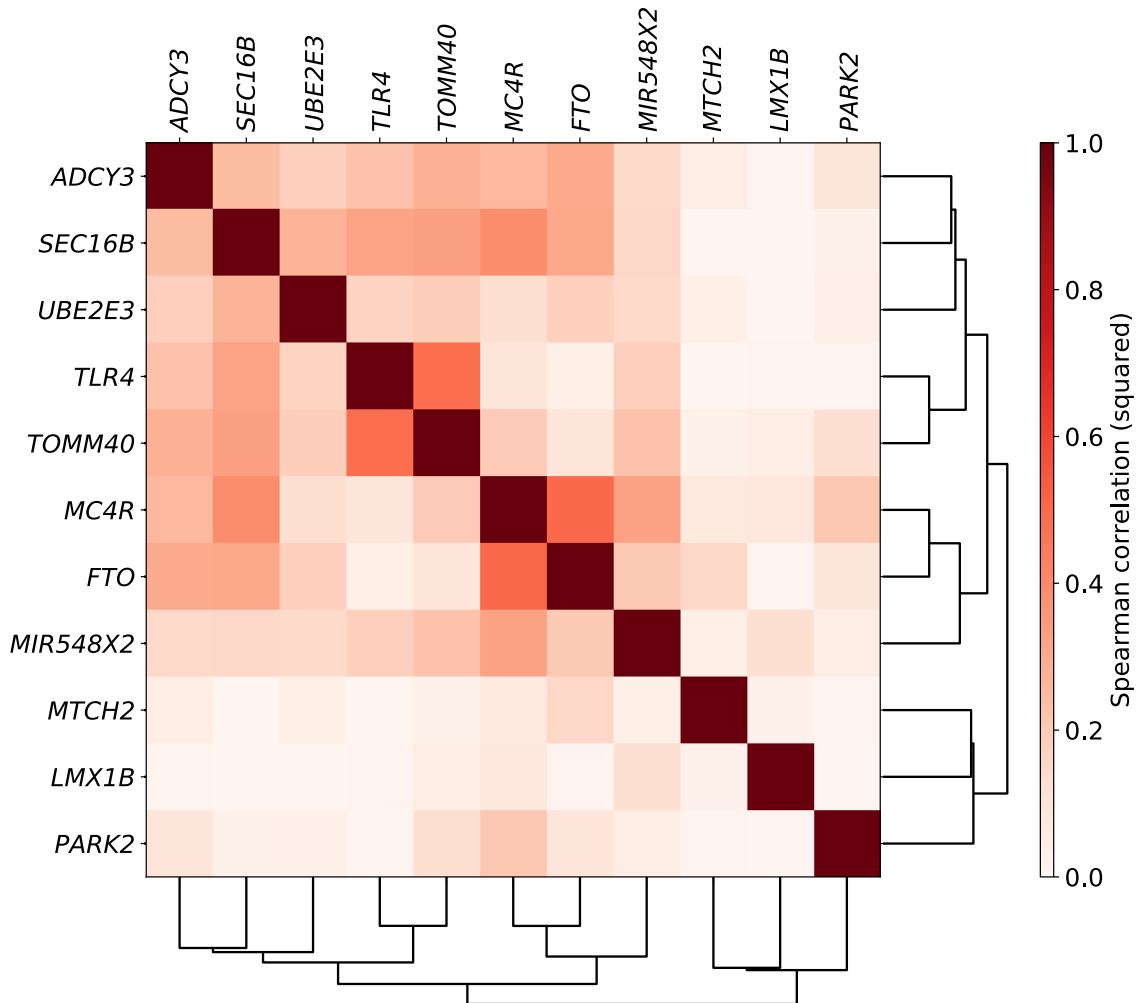


comparisons, I also compared the out-of-sample predictions and in-sample estimates of only the  $G \times E$  component of the allelic effect for the test set of individuals. This was achieved by subtracting the predicted persistent effect (added to generate the per-individual allelic effects, see Section 3.2.3) and the in-sample estimate of the variant effect using all testing individuals, respectively (Fig. 3.15d). Whilst, I note that the estimated and predicted allelic effect sizes may be overestimated due to the winner's curse<sup>360</sup>, the consistency in the direction of effect between the estimates and predictions suggests that I am able to identify individuals at the extreme ends of the risk spectrum. The out-of-sample per-individual allelic effect predictions also provide further evidence for a possible opposite direction of effect at *PARK2*.



**Fig. 3.15 Out of sample prediction of per-individual allelic effect sizes** | Assessment of per-individual allelic effects for 11 GIANT variants with evidence for G×E (Benjamini-Hochberg adjustment, FDR < 0.05), considering a 50:50 split of the cohort into training and test fractions. **(a)** Violin plot, displaying the distributions of the in-sample estimated allelic effect sizes for the training fraction (as in Fig. 3.14). Estimated persistent genetic effects are shown by the red bars and the green bars indicate the top and bottom 5% quantiles of variation in effect sizes due to G×E. P values denote the significance of genetic effects within the respective strata, assessed using LMM-Renv. Nominally significant associations ( $P < 0.05$ ) are highlighted in red. **(b)** Analogous allelic effect size distributions as in **a**, for the test fraction (out-of-sample predictions, using only the environmental states of test individuals). Yellow crosses denote the mean predicted genetic effect within the top and bottom 5% strata (considering strata with nominally significant associations in the training fraction, i.e. P values that are highlighted in red in **a**). **(c)** Scatter plot of allelic effect sizes, displaying out-of-sample predictions for the extreme strata of test set individuals (yellow crosses in **b**, x-axis) versus in-sample estimates for the extreme strata of test set individuals, obtained using LMM-Renv (using the genotypes and phenotypes of the test set individuals in the strata). Different GIANT variants are coded in colour. **(d)** Analogous scatter plot as in **c**, however displaying the differences between genetic effects in the 5% strata and population estimates of persistent effects (effect sizes due to G×E).

Finally, I examined whether it is the same environmental profiles that put individuals at increased and decreased trait risk across different loci, by calculating the squared Spearman's correlation between pairs of the 11 GIANT loci with significant  $G \times E$  effects (FDR 5%, Benjamini-Hochberg adjusted<sup>86</sup>; Fig. 3.16). This analysis suggests that there is some shared environmental burden across loci due to  $G \times E$  effects, in particular if individuals carry the risk increasing alleles at both *MC4R* and *FTO* or at both *TLR4* and *TOMM40* and again suggests that the interaction effect at *PARK2* is driven by different environmental exposures to those driving the  $G \times E$  effect at *FTO*, *MC4R* and *SEC16B*.



**Fig. 3.16 Rank correlation of per-individual genetic effect sizes across loci for UK Biobank data** | Shown are squared Spearman correlation coefficients of per-individual allelic effects estimates for 11 GIANT variants with evidence for  $G \times E$  (Benjamini-Hochberg adjustment,  $FDR < 0.05$ ).

### 3.5 Summary and discussion

In this chapter, I have shown the practical utility of StructLMM to identify variants with a  $G \times E$  component that alter BMI risk, using 64 lifestyle based environmental variables and 252,188 individuals from UK Biobank.

Initially, I focussed on testing for interactions at the 97 loci identified by the GIANT consortium<sup>316</sup>, confirming previous findings of a  $G \times E$  effect at *FTO*<sup>195–206,213,214,324</sup> and in addition replicating a  $G \times E$  effect at *MC4R*, which was previously found to interact with physical activity in women aged 20 – 40 in a cohort of Nordic individuals<sup>196</sup>. This analysis also revealed a further seven loci with significant  $G \times E$  effects (FDR 5%, Benjamini-Hochberg adjusted<sup>86</sup>; see Supplementary Table 3 in Moore *et al.*<sup>245</sup>). I also considered a genome-wide  $G \times E$  analysis, where 17,606 significantly associated BMI variants ( $P < 5 \times 10^{-8}$ , LMM-Renv) were tested for  $G \times E$ , yielding 23 loci with significant interaction effects (FDR 5%, Benjamini-Hochberg adjusted<sup>86</sup>; see Supplementary Table 3 in Moore *et al.*<sup>245</sup>).

I then applied the StructLMM joint association test, which identified 23 loci ( $P < 5 \times 10^{-8}$ , see Supplementary Table 4 in Moore *et al.*<sup>245</sup>) with  $G \times E$  contributions that were not identified by conventional association tests (Fig. 3.9 and 3.10). Fourteen of these loci remain exclusively detected by StructLMM (using a distance threshold of  $+/- 500$  kb), even when the sample size is  $\approx 2.7$  times as large (considering the recent study conducted by Yengo *et al.*<sup>317</sup>, which used 681,275 individuals, including 456,426 UK Biobank individuals) as that considered for the analyses presented in this chapter. I note that StructLMM is not intended as a replacement for conventional association tests, as it will be less powered at loci where there is virtually no environmental dependency (Fig. 3.10), since StructLMM is penalised for grid searching. Instead, it should be regarded as a tool to identify additional loci that are subject to  $G \times E$  effects.

Whilst this work demonstrates the benefits of using StructLMM to identify variants in the presence of  $G \times E$  effects, the identification of novel loci with strong  $G \times E$  (large values of  $\rho$ ) using the StructLMM joint association test not identified by standard linear models, suggests that current methods which select variants for interaction testing based on association results obtained from standard linear models are not optimal. In the next chapter, I describe an improved strategy that enables genome-wide interaction scans with sufficient power to detect  $G \times E$  after multiple testing correction. Testing for interaction effects at all genome-wide variants has the added benefit of relaxing the assumption that the lead association and lead

interaction variants are in high LD with one another.

A further increase in the number of significant loci may be achieved through using a greater number of the available UK Biobank individuals. Currently, I remove individuals with any missing environmental data but a possible extension to the existing StructLMM framework would be to incorporate environment imputation methods that leverage the correlation between the considered variables<sup>361</sup>, for example using multivariate normal models<sup>362</sup>. Such imputation methods will become increasingly important as the number of environments considered for analyses increases and thus the number of individuals removed due to missing data increases.

Finally, I showed that further exploration of loci identified by StructLMM is also possible using the methods implemented within StructLMM, as described in Section 3.2. Exploration of the environments driving the observed  $G \times E$  effects at four loci with significant interactions (FWER 5%) using Bayes factors, provides evidence that multiple environments contribute to the interaction effects and that these environments vary across the loci (Fig. 3.12 and 3.13).

Given that these interactions are likely poly-environment in nature, it is perhaps more pertinent to explore the aggregate environment effects that lead to increased or decreased trait risk. This analysis suggested that the interaction effect at *PARK2* can both increase and decrease trait risk (i. e. there is no risk increasing allele per se but that the risk increasing allele is dependent on the aggregate environmental exposure). Determining whether observed interaction effects are truly due to interactions with the environment or if instead the environments are acting as proxy for potentially multiple genetic variants, such that the identified interaction effect is actually an epistatic effect is particularly important for the interpretation and utility of interactions that display opposite directions of effect. Whilst identification of epistatic effects may reveal functional mechanisms such as genetic regulation or physical interaction between the genetic products (proteins), interactions truly due to the environment suggest that environmental changes based on the genotypes carried at such loci may be particularly influential for trait risk.

I finally assessed the correlation of the estimated aggregate environment across the 11  $G \times E$  loci (FDR 5%, Benjamini-Hochberg adjusted<sup>86</sup>), thereby exploring the impact of environmental exposures in the context of genetic risk scores. Whilst there is evidence of strong correlation at some loci, such that BMI risk is strongly exacerbated if an individual carries risk increasing alleles at multiple loci (e. g. at both *MC4R* and *FTO* or at both *TLR4* and *TOMM40*) and lies at the extreme of the environmental risk distribution, this is not the case for the majority of loci

(Fig. 3.16). This suggests that set test based approaches that aggregate  $G \times E$  effects across environments as is done in StructLMM are more appropriate than set test methods that aggregate  $G \times E$  effects across variants considering a single environmental variable (as described in Sections 2.1 and 2.2.7). Furthermore, this suggests that effective lifestyle based intervention should be individually tailored based on the genetic risk variants that an individuals carries, rather than ubiquitous across the population (that is not to say that everybody won't benefit from a healthy lifestyle). This can be thought of as a realisation of personal medicine.

Since the characterisation of loci is performed using the same dataset as for discovery, these results should be interpreted with caution and validation with an independent data set is ultimately required. However, ascertainment of a suitable dataset where the same or very similar sets of environmental data is collated, will be difficult, in particular, due the fact that environmental exposures vary with geographical location.

## Chapter 4

# Exploring genotype-environment interaction effects across different phenotypes

### 4.1 Introduction

The idea of phenome-wide association studies (PHEWAS) is not new<sup>151,154,363–366</sup>; however, due to the lack of suitable data the first proof of concept study was not conducted until 2010<sup>154</sup>. Whilst GWAS focus on a phenotype of interest, testing for associations at all available genetic variants, PHEWAS are a complementary hypothesis generating approach that is variant focussed, systematically testing a large number of phenotypes for associations<sup>151,155,367</sup>.

PHEWAS can be classified as comprehensive or targeted; comprehensive approaches test for associations at all available phenotypes, whilst targeted studies consider a limited number of traits and diseases<sup>151</sup>. The first conducted phenome-wide analysis is an example of a comprehensive study that focussed on five variants with previous disease associations, testing for associations with 733 phenotypes, using 6,005 European-American individuals (a subset of the individuals available from the electronic medical records and genomics (eMERGE) network)<sup>151,154,368</sup>. This analysis was later expanded to test for associations between 3,144 SNPs (listed in the GWAS catalog<sup>152</sup>) and 1,358 EMR-derived phenotypes in 13,835 individuals of European ancestry, revealing 63 potentially pleiotropic (when a locus affects multiple traits; see Section 1.2.4 for further details) associations ( $\text{FDR} < 0.1$ )<sup>151,369</sup>. A recent study, examined associations comprehensively for both genotypes and phenotypes,

considering a set of 635,525 genotyped SNPs and 541 binary ICD-9 codes and 25 continuous clinical laboratory measurements using 38,622 unrelated samples of European American ancestry from the DiscovEHR study (a collaboration between Geisinger and Regeneron Genetics Center)<sup>158</sup>. However, analysis at this scale did require the use of a cloud computing platform<sup>158</sup>. Examples of targeted PHEWAS, include testing for associations with different cancer types<sup>370–377</sup> and autoimmune diseases<sup>151,378</sup>.

As well as the identification of possible pleiotropic effects, revealing potentially shared biological mechanisms and thus advancing our understanding of complex traits and diseases<sup>8,139,140,156</sup>, PHEWAS can be used to aid the drug discovery process<sup>151</sup>. There is evidence to suggest that genetic associations with a disease predict potential drug targets<sup>155</sup>; for example a study conducted by Sanseau *et al.* found that as of February 2011 15.6% of genes identified through GWAS are existing drug targets compared with just 5.7% of genes when considering the whole genome<sup>155,379</sup>. Therefore, PHEWAS may reveal opportunities for drug repurposing, whereby existing drugs may be used to treat other diseases; this can be achieved by focussing on loci which are known drug targets and testing whether these loci are associated with other traits<sup>155,367,380,381</sup>. Alternatively, if opposite directions of variant effects across different traits are observed, then this may be suggestive of adverse drug reactions, such that a locus may be discounted as a viable drug target<sup>381</sup>. Finally, PHEWAS can be used to identify novel drug targets that are potentially highly profitable through the identification of loci that affect multiple phenotypes with minimal evidence of opposite directions of effect<sup>381</sup>.

Tools and software that facilitate easy running of such phenome-wide association scans are starting to be developed<sup>151</sup>. In particular, automated data harmonisation tools that encompass QC procedures to deal with diverse phenotypic data, including the presence of both binary and continuous phenotypes, are emerging<sup>151,382</sup>.

PHEASANT is one such recently developed tool, designed to perform automated phenome scans in UK Biobank<sup>382</sup>. In addition databases are emerging, including PhenoScanner (<http://www.phenoscanter.medschl.cam.ac.uk>) that collates results irrespective of P value significance from large scale association scans, enabling rapid phenome-wide scans<sup>383</sup> and the PHEWAS catalog (<https://phewascatalog.org>)<sup>369</sup> that stores nominally significant results ( $P < 0.05$ ) from phenome-wide studies that exceed a given threshold, developed in a similar vein to the GWAS catalog<sup>152</sup>.

However, despite these developments, the significance threshold for PHEWAS is



not well established. Bonferroni multiple testing is likely too stringent due to the correlation between phenotypes<sup>367</sup> (see Section 1.1.6). Whilst PC based methods to calculate the effective number of traits (described in Section 1.1.6) can be applied if the considered genetic variants are independent of one another, calculation of the effective number of tests is not straightforward when many variants are considered due to the potential correlations between both the phenotypes and the variants<sup>384</sup>. Furthermore, a proportion of the tests represent already known genotype-phenotype associations and can therefore be considered as positive controls rather than experimental tests of association<sup>384</sup>. Differences in the power to detect associations across different traits due to differing sample sizes, in particular for binary phenotypes where the number of cases can vary substantially, is a further complication that is yet to be addressed<sup>158</sup>. Current standard practice is to therefore, consider both a lenient 10% FDR threshold<sup>369,381,385</sup> (or less frequently a Bonferroni multiple testing correction based on the effective number of tests as described in Section 1.1.6), e. g. Verma *et al.*<sup>158</sup>) applied to all variant-phenotype associations tested for and a 5% Bonferroni or FDR threshold for each analysed variant (i. e. applying a multiple testing correction only across considered phenotypes)<sup>158,369,386</sup>.

It is well known that complex traits and diseases are influenced not only by genetic factors, but in addition by environmental exposures and interaction effects between genetic factors and the environment. However, all phenome-wide studies to date, except for one, do not consider interaction effects. The exception is a pilot phenome-wide epistasis scan (which the authors refer to as a phenome-wide interaction study, PHEWIS for short), that considers 26 phenotypes obtained from 2,547 treatment-naive AIDS patients<sup>387</sup>. Epistatic interactions were tested for at 1,773 variants that replicated in marginal association studies and lie in functionally active regions based on chromatin state annotations, such that in total 40,842,828 interaction tests were conducted<sup>387</sup>. 12,146 significant SNP-SNP interactions (FDR adjusted  $P < 0.01$ ) across only two of the phenotypes were identified (details of the overlap between the two traits were not provided)<sup>387</sup>.

PHEWIS may substantially increase our knowledge of the contribution of interaction effects to complex traits and diseases. They may reveal that interaction factors are trait or locus specific, or reveal a greater overlap in the loci that impact different traits than that observed when considering only persistent genetic effects.

Furthermore, PHEWIS can be used to explore the relative importance of  $G \times E$  effects compared to persistent genetic effects across different traits by evaluating the amount of phenotypic variation explained by persistent and  $G \times E$  effects (see Section 3.2.1) at different loci or in aggregate across multiple loci. The importance

of different factors for explaining phenotypic variability is a fundamental question in the field<sup>44,388</sup>.

Typically, phenotypic variation is partitioned into genetic and non-genetic components, with the fraction of phenotypic variation explained by genetic factors referred to as heritability<sup>44,45</sup>. Heritability, is a population based measure, since by definition it is dependent on the importance of non-genetic factors which can vary across samples<sup>44,388</sup>.

There are two commonly used definitions of heritability, narrow and broad sense heritability. Narrow sense heritability is the amount of phenotypic variation that is explained by additive persistent genetic effects (such that the non-genetic component described above does in fact contain non-additive genetic effects) whilst broad sense heritability is the amount of phenotypic variation explained by genetics, including dominant and interaction effects<sup>389</sup>.

Twin and adoption studies are classic designs for heritability studies, that provide broad sense heritability estimates on the assumption that the amount of shared environmental exposures can be accounted for<sup>389,390</sup>. By comparison, for the classical family-based design or more recent population based heritability studies, differences in environmental exposures across individuals cannot be controlled and thus these analyses estimate narrow sense heritability<sup>389</sup>.

Initial, population based heritability estimates were calculated using variants that were deemed significant based on GWAS, resulting in very low heritability estimates<sup>389</sup>. There has since been a shift to estimating heritability based on all measured genetic variants, increasing population based heritability estimates<sup>389</sup>. Despite this increase, classic heritability estimates tend to be higher than population based heritability estimates, with the difference described as ‘missing heritability’<sup>389,391</sup>. There is a consensus that some of this difference is attributable to epistasis and  $G \times E$ <sup>88,389–391</sup>, although inclusion of these additional phenotypic partitions and subsequent estimation of phenotypic variation explained by interaction effects is not typical.

There are two studies which are notable exceptions that attempt to estimate the amount of phenotypic variation explained by interaction effects. The first was a small-scale (820 individuals) family-based study that examined the contribution of  $G \times E$  effects for four type 2 diabetes related traits, for each of 15 dietary and lifestyle environmental factors<sup>392</sup>. The phenotypic variation explained by  $G \times E$  effects was estimated using SNPs genome-wide, identifying zero, four, two and five environments

for fasting glucose, fasting insulin, HOMA-IR and HOMA-B that explain more phenotypic variation than persistent genetic effects<sup>392</sup>. However, the family-based design is a key weakness of this study, leading to possible overestimation of both persistent and  $G \times E$  contributions and bias in the variance explained due to greater sharing of environments within than between families<sup>392</sup>. Furthermore, environmental contributions were estimated separately for each exposure such that the summed phenotypic variation explained by  $G \times E$  across all environments was greater than 100% due to correlation between the environments<sup>392</sup>. The second study primarily used phenotypic variation analyses to determine the presence of interaction effects on BMI for age and eight other dietary and lifestyle based factors (using 97,510 individuals from UK Biobank), concluding that the majority of genotype-environment interactions explain very little phenotypic variation, with the exception of age and smoking which are estimated to explain 8.1% and 4.0%<sup>388</sup>.

In this work, I perform an initial exploration of interaction effects across different phenotypes in a similar manner to existing targeted PHEWAS and examine the contribution of  $G \times E$  effects jointly at multiple environments across different phenotypes, relative to persistent genetic effects. Hypothesising that variants with largely persistent genetic effects for one trait might also influence other traits (in particular those that are correlated with the primary trait) in combination with environmental exposures, I use StructLMM to test for interaction effects between 64 lifestyle based factors and variants that are significantly associated with basal metabolic rate (BMR) on nine additional phenotypes, body fat percentage (body fat %), weight, BMI, diastolic blood pressure (DBP), systolic blood pressure (SBP), hip circumference (HC), waist circumference (WC), standing height (height) and peak expiratory flow (PEF). BMR is the amount of energy a body needs at rest for maintenance, and in UK Biobank participants it was calculated based on body composition through the use of bioelectrical impedance analysis.

To address the potentially large number of variants significantly associated with BMR that are to be taken forward for  $G \times E$  testing, combined with the relatively low power of interaction tests, I employ the use of a conditional FDR (cFDR) multiple testing correction<sup>393,394</sup>. The cFDR was originally developed for application to two independent association scans, leveraging genetic pleiotropy between the traits<sup>394</sup>. Application of this method to identify variants associated with schizophrenia conditional on bipolar disorder increased the number of loci (LD,  $r^2 < 0.2$ ) from six using a uniform 5% FDR to 58 using a conditional 5% FDR<sup>394</sup>. This work was later extended, such that the correction could be applied to experimental designs with common controls, in which the two association scans are not completely independent

of one another<sup>393</sup>.

The key concept of this multiple testing correction, is that the prior probability of a test being false is not uniform across all hypotheses tested and as a result the use of a uniform FDR is not optimal for maximising the number of discoveries for a given number of expected false discoveries<sup>395</sup>. This prior probability is described through the use of a covariate, which for the aforementioned studies are the association P values obtained from conducting an association scan for one trait that can be used as prior information on the likelihood that a variant is associated with the second trait. Similar methods are also available that use other covariates, such as functional annotations that pre-determine categories of variants, with some categories of variants more likely to be enriched for associations. In these categorical settings, the FDR is applied separately to each category, known as the stratified FDR<sup>396,397</sup> or the independent hypothesis weighting method<sup>395</sup>. Since the cFDR, effectively assigns each variant to its own category, it can be thought of as a continuous version of these stratification based methods<sup>85</sup>.

In this chapter, I show that it is possible to condition the interaction results obtained using StructLMM on the association results obtained from marginal association analyses using LMMs. In this setting, the cFDR can be interpreted as bridging two analysis designs. The first that is commonly used, filters variants for interaction testing based on marginal association results<sup>175,177,180,193,219–231</sup>, thereby only enabling detection of interaction effects at variants with strong marginal priors. The second is a genome-wide interaction scan that tests all available variants, which thus suffers from a high multiple testing burden. Explicitly, in Section 4.2, I provide mathematical details of the cFDR and the conditions that need to be satisfied and in Section 4.3.4, I demonstrate its potential to identify loci with significant interaction effects.

## 4.2 Methods

In this section, I will describe methods that I use throughout this chapter. I will start by describing the data pre-processing steps that were applied to the UK Biobank data. Whilst very similar to the data pre-processing steps described in Section 3.3.1, a major difference in this work is the construction of a pipeline that minimises the per phenotype QC, similar to the advantages of previous automated pipelines such as PHEASANT<sup>382</sup>. This will be particularly important in future work when a greater number of phenotypes are considered. This will be followed by details of

the calibration assessment methods and the method used to identify independent loci. I will then describe the cFDR multiple testing correction and finally, the method used to estimate the amount of phenotypic variation explained by different components.

### 4.2.1 UK Biobank data preprocessing

The analyses in this chapter have been conducted using the full release of UK Biobank (Application 41672)<sup>215,271</sup>.

#### Phenotype and environment pre-processing

First, I identify individuals that are to be excluded from all analyses. These are poor quality samples flagged by UK Biobank using the field ‘het.missing.outliers’ (based on heterozygosity and the amount of missing data), individuals with more than ten 3<sup>rd</sup> degree relatives using the field ‘excess.relatives’ in the released ‘Sample-QC’ file and those that withdrew consent (104 in total). I then kept only those individuals that were genetically ‘White British’ based on the field ‘in.white.British.ancestry.subset’ and finally removed any remaining individuals that were listed in field ‘ID1’ using the released ‘Relatedness’ file such that there were no relatives (3<sup>rd</sup> degree or closer) included in the analysis. This left a total of 335,587 individuals.

BMR, body fat %, weight, BMI, DBP, SBP, HC, WC, standing height and PEF phenotype data are all based on ‘Instance 0’ of UK Biobank fields ‘23105’, ‘23099’, ‘21002’, ‘21001’, ‘4079’, ‘4080’, ‘49’, ‘48’, ‘50’ and ‘3064’, respectively (see <http://biobank.ctsu.ox.ac.uk/crystal/> for details).

The same set of lifestyle based factors used in Chapter 3 (see Section 3.3.1) were selected. Namely, 20 environmental variables: 9 ordinal dietary variables (‘Oily fish intake’, ‘Non-oily fish intake’, ‘Processed meat intake’, ‘Poultry intake’, ‘Beef intake’, ‘Lamb/mutton intake’, ‘Pork intake’, ‘Cheese intake’ and ‘Salt added to food’), three continuous dietary variables (‘Cooked vegetable intake’, ‘Bread intake’, ‘Tea intake’), three physical activity variables (‘Number of days/week walked 10+ minutes’, ‘Number of days/week of moderate physical activity 10+ minutes’, ‘Number of days/week of vigorous physical activity 10+ minutes’), ‘Alcohol intake frequency’, ‘Sleep duration’, ‘Townsend deprivation index’, ‘Smoking status’ and ‘Time spent watching television’, using the data from ‘Instance 0’ (see <http://biobank.ctsu.ox.ac.uk/crystal/> for details), were selected. Again following

the pre-processing steps conducted in Chapter 3 (see Section 3.3.1), for the three continuous dietary variables, I set values exceeding the 99<sup>th</sup> percentile (based on the 335,587 individuals) to missing. For ‘Sleep duration’, I set the top and bottom percentiles (based on the 335,587 individuals) to missing and for each individual calculated the squared deviations from the mean sleep duration, creating an additional environmental variable, ‘Squared sleep duration res.’ (21<sup>st</sup> environmental variable). For ‘Time spent watching television’, less than 0.5 hours of was encoded as 0.5 and I set values in the upper and lower percentile (based on the 335,587 individuals) to missing. A total of 251,720 individuals have no missing environmental data.

Individuals with data missing for any of the environmental variables or the phenotype under study were removed as part of the testing procedure, such that a data set for each considered phenotype after removing individuals with any missing data does not require storage. A total of 248,015 (BMR), 247,904 (body fat %), 251,261 (weight), 251,205 (BMI), 235,626 (DBP), 235,623 (SBP), 251,441 (HC), 251,443 (WC), 251,390 (height) and 231,211 (PEF) individuals were included in the analyses. A rank-based inverse normal transformation (commonly used for genetic analyses, see Pain *et. al*<sup>398</sup>) was then applied to the phenotype under study.

## Generation of principle components

Genetic principle components (PCs) were generated using FlashPCA version 2.0<sup>339</sup>, based on the set of 147,604 SNPs flagged by the field ‘used.in.pca.calculation’ in the released ‘Sample-QC’ file and the 335,587 individuals that passed the sample QC procedures and did not have any missing environmental data (described above). Ten PCs were used to control for population structure.

## Variant selection

Variants selected for exploration across different phenotypes, were those that exceeded the genome-wide significance threshold of  $5 \times 10^{-8}$  based on publicly available summary statistics (available for download from <http://www.nealelab.is/blog/2017/9/11/details-and-considerations-of-the-uk-biobank-gwas>, downloaded on March 1st 2018)<sup>399</sup> from an association scan of BMR.

This previously conducted GWAS was based on 337,199 samples and 10,894,597 variants, including ten PCs and sex as covariates in a linear regression analysis.

Briefly, sample QC involved keeping only unrelated individuals of ‘White British’ ancestry and variants based on the HRC panel with an INFO score  $> 0.8$ , MAF  $> 0.1\%$  and HWE  $P > 1 \times 10^{-10}$  (see <http://www.nealelab.is/blog/2017/9/11/details-and-considerations-of-the-uk-biobank-gwas> for further details).

## Environment covariance, $\Sigma$ , and model covariates

The environmental covariance matrix was generated identically to that described in Chapter 3 (see Section 3.3.1 for details) for the set of individuals included in each phenotype analysis. The covariates included in the analyses were also identical to those described in Chapter 3 (see Section 3.3.1 for details).

### 4.2.2 Calibration

Calibration was assessed by permuting the 173,297 variants (over the individuals included for each phenotype) on chromosome 20 such that any true association and interaction signals should be destroyed. The resulting 173,297  $P$  values from the StructLMM interaction test (StructLMM-int) or association test (LMM-Renv) were used to generate QQ plots of the expected negative log  $P$  values (x-axis) versus the observed negative log  $P$  values (y-axis) and inflation parameters,  $\lambda_{GC} = \frac{\log_{10}(m)}{\log_{10}(0.5)}$ , where  $m$  is the median  $P$  value of the 173,297 variants tested, were calculated.

### 4.2.3 Defining loci

Independent loci were defined based on the variants significantly associated ( $P < 5 \times 10^{-8}$ ) with BMR using the publicly available summary statistics (see Section 4.2.1 for further details) by iteratively (i) selecting the most significant variant and (ii) clumping all variants in LD,  $r^2 \geq 0.1$  (calculating LD using 10,000 UK Biobank individuals randomly selected from the 251,720 individuals that passed sample QC and had no missing environmental data) within  $+/- 1$  Mb, until no variant was left. This resulted in 1,104 loci and these defined loci (and the variants assigned to each locus) are kept constant throughout all analyses.

#### 4.2.4 cFDR

The cFDR adjusted P values are calculated as follows. Let  $p_m^P$  and  $p_m^C$  be the P values of the  $m^{\text{th}}$  variant for the principal trait of interest and the second trait which is to be used as the covariate conditioned on, respectively. Then the P values for the principal trait at all  $M$  tested variants are given by  $\{p_1^P, p_2^P, \dots, p_M^P\}$  and the P values at all  $M$  tested variants for the second trait (to be conditioned on) are given by  $\{p_1^C, p_2^C, \dots, p_M^C\}$ . Then the cFDR adjusted P value for the  $m^{\text{th}}$  variant, is given by:

$$p_{m_{\text{adj}}}^P = \frac{p_m^P}{\text{Number pairs } (p_i^P, p_i^C) \text{ with } p_i^P \leq p_m^P \text{ and } p_i^C \leq p_m^C} \times (\text{Number } p_i^C \leq p_m^C). \quad (4.1)$$

The  $m^{\text{th}}$  variant is declared to be significantly associated with the principal trait of interest if  $p_{m_{\text{adj}}}^P < \alpha$  (often  $\alpha = 0.05$ ).

A conventional hard filtering approach tests only variants with  $p^C < \text{global threshold}$  (e. g. global threshold =  $5 \times 10^{-8}$ ) for association at a second trait, with a uniform FDR applied to all selected variants. In comparison, from Eq. 4.1, it can be seen that with the cFDR approach the threshold for selecting variants is successively relaxed until all variants are tested for association at a second trait, requiring increasing evidence of association with the second trait, for a significant effect to be declared. As a result the cFDR can be viewed as a ‘soft’ thresholding approach, yielding a greater number of identified true associations if associations are shared across the two traits<sup>393</sup>.

Conditions for the cFDR to hold are that the principal and covariate tests are independent of one another and that both tests are calibrated under the null.

In this work, I define the interaction P values obtained from the StructLMM interaction test as those coming from the principal trait and the association P values obtained from LMM-Renv as the covariate to condition on. Dai *et. al*<sup>229</sup> provide a formal proof that test statistics (using the LRT and score test) based on any two nested models are independent, which without loss of generality holds for the StructLMM interaction test and LMM-Renv. Explicitly, the tested term in StructLMM interaction test,  $\mathbf{x} \odot \boldsymbol{\beta}_{G \times E}$  (see Eq. 2.9) is the only term not included in LMM-Renv model (see Eq. 2.121), meaning that the models are nested and as a result the test statistics corresponding to the marginal genetic and interaction effects are independent of one another. I also show that the StructLMM interaction and the LMM-Renv association test are calibrated for all considered phenotypes in



Section 4.3.2.

Analogous to the aforementioned application to independent association scans, this conditional multiple testing correction in the interaction test setting will increase the number of true interactions identified, if interaction signals are enriched for amongst the association signals (here association signals means evidence for associations and does not require associations to be deemed significant).

### 4.2.5 Estimation of phenotypic variance explained

For each locus (see Section 4.2.3 for definition) I select the lead variant per phenotype for both the association test (smallest P value obtained using LMM-Renv) and the interaction test (smallest cFDR adjusted P values obtained using StructLMM-int). For each selected variant (in total,  $2 \times 1,104 \times 9 = 19,872$ ), I estimate the fraction of the phenotypic variance explained by marginal genetic (G) and interaction (G×E) effects as described in Section 3.2.1. I then sum the marginal genetic contributions based on the lead association variants and the G×E contributions based on the lead interaction variants over the loci, to give an aggregate estimate of the phenotypic variation explained by marginal genetic and interaction effects at the defined 1,104 loci (see Section 4.2.3). This approach is analogous to that used to estimate the heritability explained by significantly associated genetic loci obtained from GWAS<sup>389,400</sup>. An estimate of the total phenotypic variation explained by both persistent and interaction effects at the considered set of 1,104 loci and 64 environments can be obtained by summing the respective aggregate estimates.

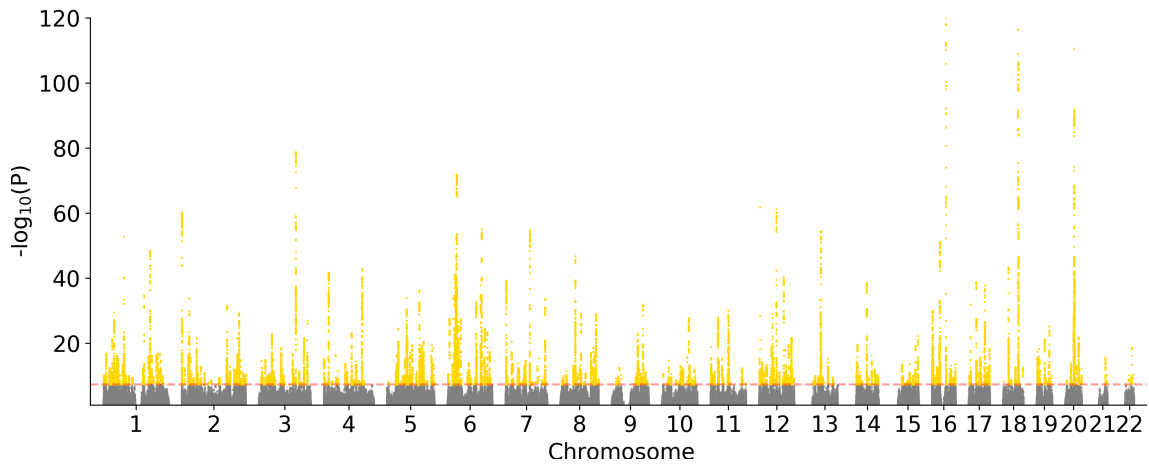
## 4.3 Results

In this section, I first outline the rationale for selecting variants associated with BMR to take forward for further analysis. I then show calibration results for the ten phenotypes considered in this analysis. This is followed by the results from association and interaction testing at the selected variants for the ten phenotypes considered in this work. Finally, I explore the relative importance of G×E effects compared to persistent genetic effects for the considered phenotypes.

### 4.3.1 Basal metabolic rate significantly associated variants

Working under the hypothesis that variants significantly associated with one trait might also influence other traits (in particular those correlated with the primary trait) in combination with environmental exposures and that traits with a relatively high estimated heritability will be good candidates to identify many such variants, I initially focus on basal metabolic rate (BMR). BMR has an estimated heritability of 28.64%, which is higher than other cardiometabolic traits, including weight (26.66%), BMI (24.63%), HC (22.30%), body fat % (22.02%), WC (20.36%), DBP (14.37%) and SBP (13.36%; all heritability estimates are based on the partitioned heritability estimates provided at [https://nealelab.github.io/UKBB\\_ldsc/h2\\_browser.urlhtml](https://nealelab.github.io/UKBB_ldsc/h2_browser.urlhtml)). I also consider height (estimated heritability, 46.23%) and PEF (estimated heritability, 9.35%) for which I don't expect to see interaction effects between the selected 64 lifestyle based environments and variants that are associated with BMR and thus these traits act as negative controls.

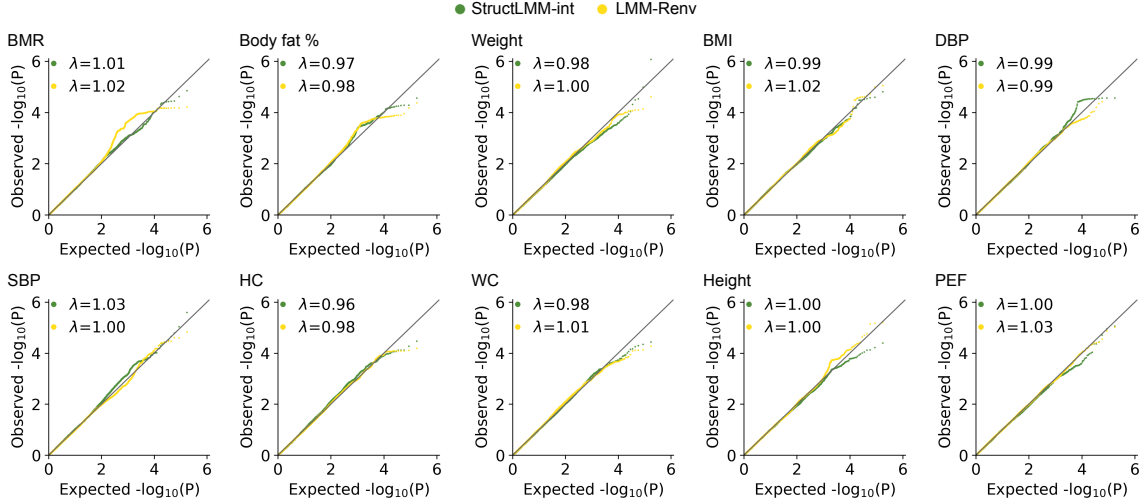
There are 57,328 variants significantly associated ( $P < 5 \times 10^{-8}$ ) with BMR (Fig. 4.1; see Section 4.2.1 for further details). These significant variants were used to define 1,104 loci ( $\pm 1$  Mb, LD  $r^2 < 0.1$ ; see Section 4.2.3 for details) used throughout the subsequent analyses for ease of comparison.



**Fig. 4.1** Manhattan plot of genome-wide association results for BMR | Manhattan plot of negative log P values (y-axis) against chromosome and position (x-axis) based on publicly available summary statistics from a genome-wide association scan of BMR using UK Biobank data (see Section 4.2.1 for details). The genome-wide significance threshold of  $5 \times 10^{-8}$  is represented by the red dashed line and variants that exceed this threshold are coloured in yellow.

### 4.3.2 Calibration assessment

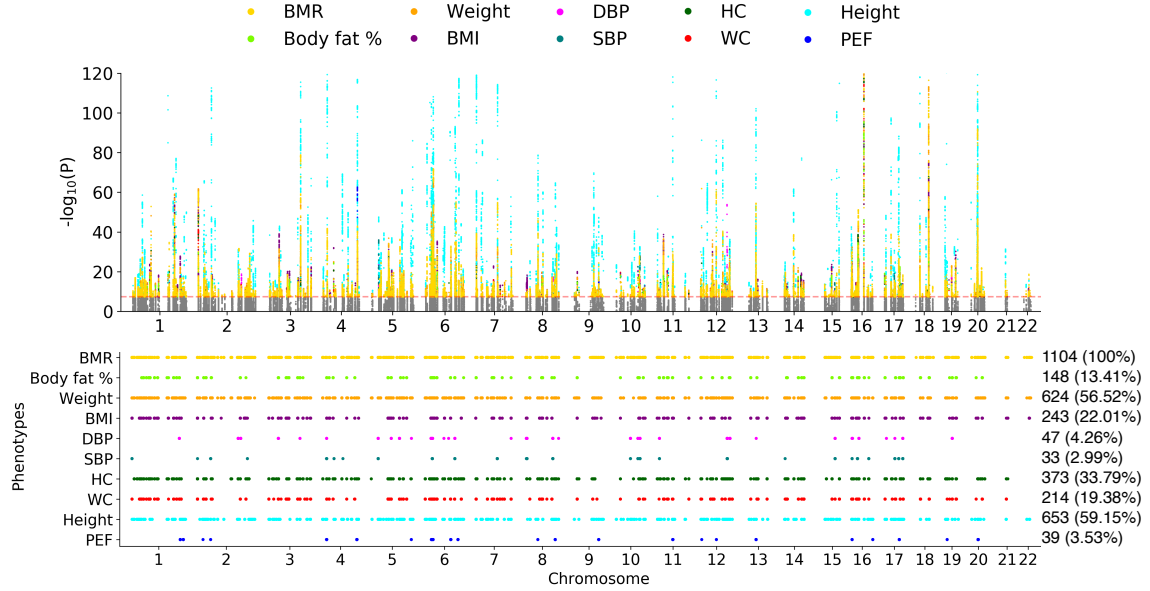
To check that the data was pre-processed appropriately and in addition to check that the data was appropriately calibrated satisfying one of the conditions required for application of the cFDR (see Section 4.2.4), I assessed the empirical calibration of LMM-Renv and the StructLMM interaction (StructLMM-int) tests (see Section 2.3 for model details) for the ten considered phenotypes. Both methods were sufficiently calibrated across all phenotypes (Fig. 4.2).



**Fig. 4.2 Calibration of interaction and association tests on UK Biobank data** | QQ plots of negative log P values from StructLMM interaction (StructLMM-int) test (green) and LMM-Renv association test (yellow) applied to UK Biobank phenotype data based on permuted genetic variants (chromosome 20). From left to right and top to bottom: basal metabolic rate (BMR), body fat percentage (body fat %), weight, body mass index (BMI), diastolic blood pressure (DBP), systolic blood pressure (SBP), hip circumference (HC), waist circumference (WC), standing height (height) and peak expiratory flow (PEF).

### 4.3.3 Association of variants with the considered phenotypes

Considering, now only the 57,328 variants that were significantly associated with BMR (as described in Section 4.3.1), I test for associations using LMM-Renv (see Section 2.3 for model details) with each of the ten phenotypes. Using the same genome-wide significance threshold ( $5 \times 10^{-8}$ ) and loci defined based on the BMR summary statistics (see Sections 4.2.3 and 4.3.1), I examine the number of the 1,104 loci that are significant associations with each of the considered traits (Fig. 4.3). The percentage of loci that are associations with other traits ranges from 2.99% for SBP to 59.15% for height.



**Fig. 4.3 Association results with the considered phenotypes** | Top: Manhattan plot of negative log P values (y-axis; using LMM-Renv) against chromosome and position (x-axis) for the 57,328 variants significantly associated ( $P < 5 \times 10^{-8}$ ) with BMR for ten phenotypes using UK Biobank data. Variants that exceed the genome-wide significance threshold ( $P < 5 \times 10^{-8}$ ) are coloured according to the phenotypes that they are associated with, with non significant results represented by grey dots. Bottom: Indicator plot with phenotypes (y-axis) against chromosome and position (x-axis), where each dot represents a variant that is significantly associated ( $P < 5 \times 10^{-8}$ ) with the phenotype. The number of loci (out of a total of 1,104) that are significantly associated with each phenotype are displayed down the right hand side with the corresponding percentage displayed in brackets. Significantly associated BMR, body fat %, weight, BMI, DBP, SBP, HC, WC, height and PEF variants are represented by yellow, light green, orange, purple, magenta, teal, dark green, red, cyan and blue coloured dots, respectively.

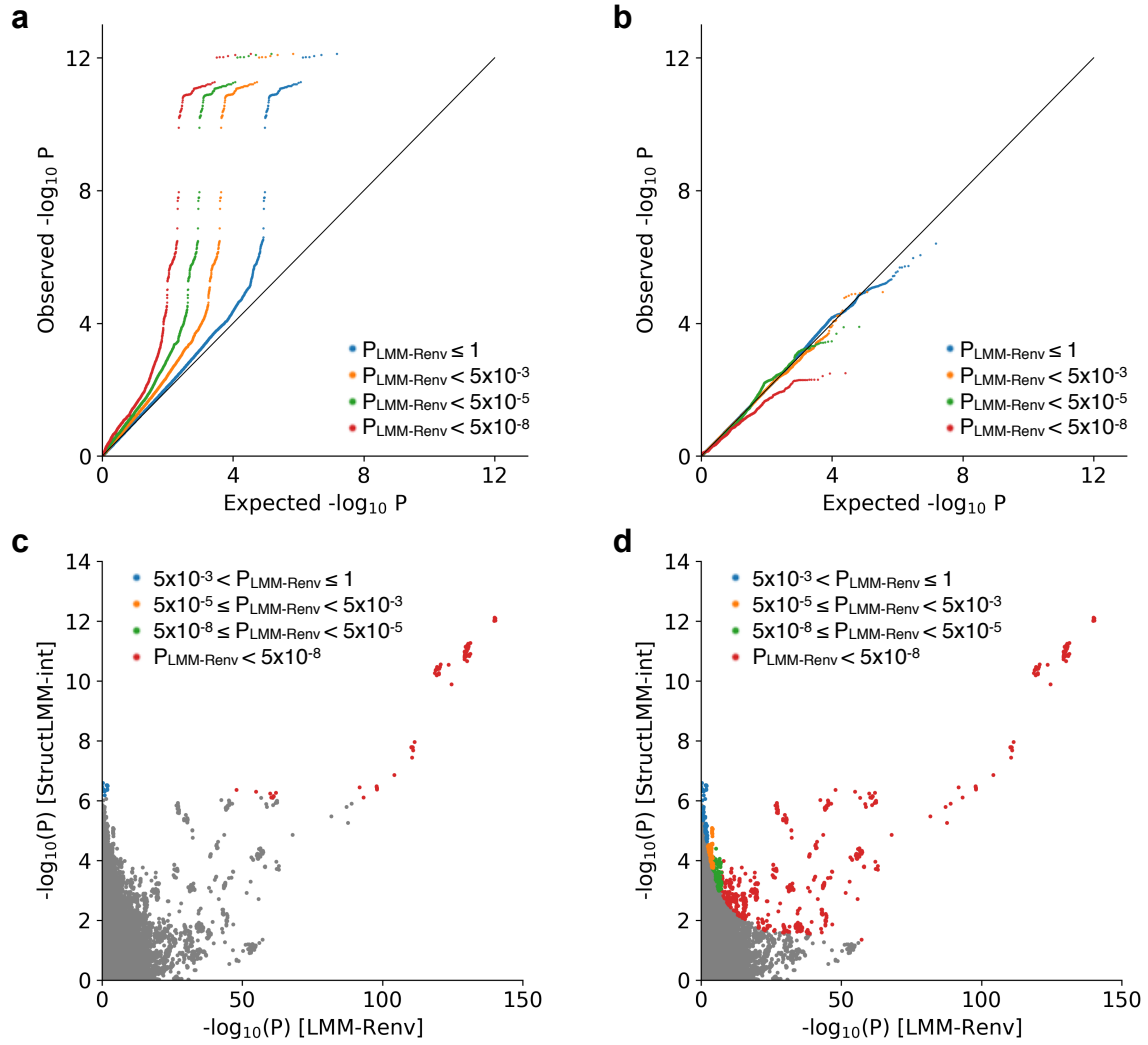
### 4.3.4 Interaction of variants for different considered phenotypes

I test for interaction effects between each of the 57,328 variants that were significantly associated with BMR (as described in Section 4.3.1) and 64 lifestyle based environments using the StructLMM interaction test (see Section 4.2.1 for details). Since interaction tests are often underpowered (see Section 1.3.2), I combine the StructLMM interaction test with a cFDR multiple testing correction (see Section 4.2.4 for a description of this method) in an attempt to increase the number of true interaction variants identified, whilst maintaining the expected number of false discoveries. I will first show the potential of this multiple testing correction method, followed by the interaction results for the different considered phenotypes.

#### Application of the cFDR for identification of $G \times E$ effects

To demonstrate the potential of the cFDR in the interaction test setting, I apply the method to genome-wide interaction results (7,515,856 SNPs as described in Section 3.3.1). As outlined in Section 1.1.6, QQ plots can be used to visualise departure of the test statistics from the null distribution, with departure in the tail, evidence that associations are present. In this setting, where I condition interaction results based on association results, I can determine if using association results as a covariate is informative by plotting stratified QQ plots. A leftward shift of the QQ plots for increasingly stringent association P value thresholds indicates an enrichment of interaction signals amongst variants that display evidence of association<sup>394</sup>. Whilst the stratified QQ plots for BMI indicate that this is the case (Fig. 4.4a), the corresponding plot for height displays no enrichment (Fig. 4.4b). The lack of enrichment when considering height also indicates that false interaction signals are not induced via this conditioning mechanism, further confirming that the association test LMM-Renv and the StructLMM interaction test are independent of one another, a condition required for the validity of this multiple testing correction method.

The number of significant interaction variants for BMI increases from 119 using a 5% FDR (Fig. 4.4c) to 1,088 using a 5% cFDR (Fig. 4.4d), corresponding to eight and 60 loci ( $+/-1$  Mb,  $r^2 < 0.1$ ). Note that the cFDR multiple testing correction identifies a superset of the variants identified using the FDR multiple testing correction, such that the cFDR does not restrict the identification of interaction variants to those with evidence of association. However, many additional variants that display some



**Fig. 4.4 Application of the cFDR to genome-wide interaction results** | (a-b) QQ plots of negative log P values from the StructLMM interaction test stratified according to the association test result obtained using LMM-Renv based on genome-wide scans for (a) BMI and (b) height. The interaction results for all variants genome-wide are shown in blue, whilst those with an association  $P < 5 \times 10^{-3}$ ,  $P < 5 \times 10^{-5}$  and  $P < 5 \times 10^{-8}$  are displayed in orange, green and red, respectively. (c-d) Scatter plots of negative log P values from StructLMM interaction test (y-axis) against negative log P values from the LMM-Renv association test (x-axis), with non-significant interaction variants shown in grey and significant interaction variants coloured according to the corresponding association P value bin using (c) a 5% FDR threshold (no variants with significant interaction effects have association P values in the range  $5 \times 10^{-8}$  to  $5 \times 10^{-3}$ ) and (d) a 5% cFDR threshold.

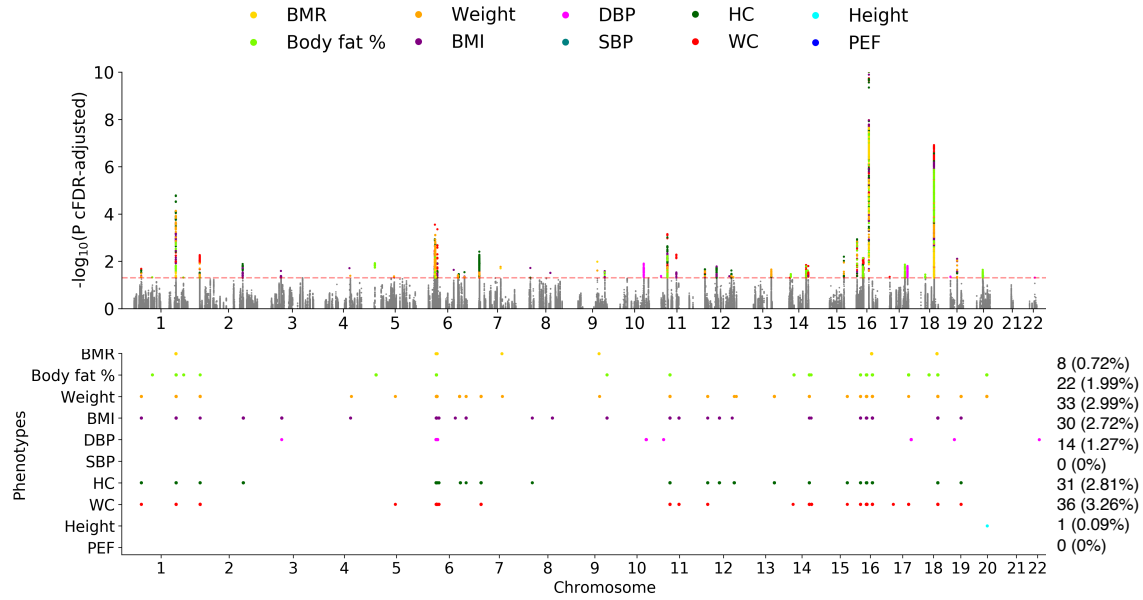
evidence of association are identified, in agreement with the theory that the presence of variants that influence trait risk may be masked by interaction effects<sup>177,188</sup>.

## Interaction results

Testing for interaction effects at the 57,328 variants (corresponding to 1,104 loci) that were significantly associated with BMR combined with a cFDR multiple testing correction reveals that as expected, the two phenotypes height and PEF, chosen as negative controls, are flat in the interaction space (with the exception of one variant and correspondingly one locus associated with height; Fig. 4.5).

In addition, as hypothesised with the exception of SBP, the number of loci with significant interaction effects is higher for other cardiometabolic traits than BMR, with only eight loci significant for BMR compared to a range of 14 - 36 loci for the other six cardiometabolic traits considered (Fig. 4.5). A total of 74 distinct loci with interaction effects are identified (see Table B1, Appendix B for a table containing results for these 74 loci). 32.43% of these loci are shared across at least three phenotypes, with the set of shared phenotypes varying across the loci (see Table B2, Appendix B for a summary of the number of loci shared for different combinations of traits). It is also worth noting, that many of the identified loci with significant interaction effects ( $P$  cFDR-adjusted  $< 0.05$ ) were not significant associations ( $P < 5 \times 10^{-8}$ ) for the same trait, again highlighting that the cFDR adjustment does not lead to the identification of interaction effects only at well established association loci. Explicitly, ten of 36 (HC), four of 27 (WC), six of 14 (DBP), three of 30 (BMI) and eight of 22 (body fat %) loci identified ( $P$  cFDR-adjusted  $< 0.05$ ) were not associated with the same trait, such that as many as 42.86% of the loci identified through interaction testing were not significant associations under the assumption of a persistent genetic effect across all individuals for DBP.

One locus with significant interaction effects lies in the *IGF1R* gene and is associated with BMR ( $P = 5.12 \times 10^{-23}$ ), weight ( $P = 5.03 \times 10^{-10}$ ) and height ( $P = 3.47 \times 10^{-68}$ ) and is found to interact with significant effects on weight ( $P$  cFDR-adjusted = 0.0113), WC ( $P$  cFDR-adjusted = 0.0462) and HC ( $P$  cFDR-adjusted = 0.00640) with *rs116222218* being the lead interaction variant for all three traits (see Table B1, Appendix B, locus number 56). In addition to its well known association with height<sup>401,402</sup>, *IGF1R* is also associated with fasting plasma glucose<sup>403</sup>. Evidence from animal studies also shows that *IGF1R* plays a role in metabolism<sup>404</sup> and in particular, that *IGF1R* gene knockout mice display evidence of altered fat mass, which in some cases was observed to be age, sex and diet dependent<sup>405-409</sup>, supporting



**Fig. 4.5 Interaction results with the considered phenotypes** | Top: Manhattan plot of negative log cFDR adjusted interaction P values (y-axis; using StructLMM-int) against chromosome and position (x-axis) for the 57,328 variants significantly associated ( $P < 5 \times 10^{-8}$ ) with BMR for ten phenotypes using UK Biobank data. Variants that exceed the genome-wide significance threshold ( $P \text{ cFDR-adjusted} < 0.05$ ) are coloured according to the phenotypes that they are interactions for, with non significant results represented by grey dots. Bottom: Indicator plot with phenotypes (y-axis) against chromosome and position (x-axis), where each dot represents a variant that has a significant  $G \times E$  effect ( $P \text{ cFDR-adjusted} < 0.05$ ) on a phenotype. The number of loci (out of a total of 1,104) that have significant interaction effects with each phenotype, based on the 64 lifestyle based environments considered, are displayed down the right hand side with the corresponding percentage displayed in brackets. Variants with significant interaction effects on BMR, body fat %, weight, BMI, DBP, SBP, HC, WC, height and PEF are represented by yellow, light green, orange, purple, magenta, teal, dark green, red, cyan and blue coloured dots, respectively.



the presence of  $G \times E$  effects at this locus.

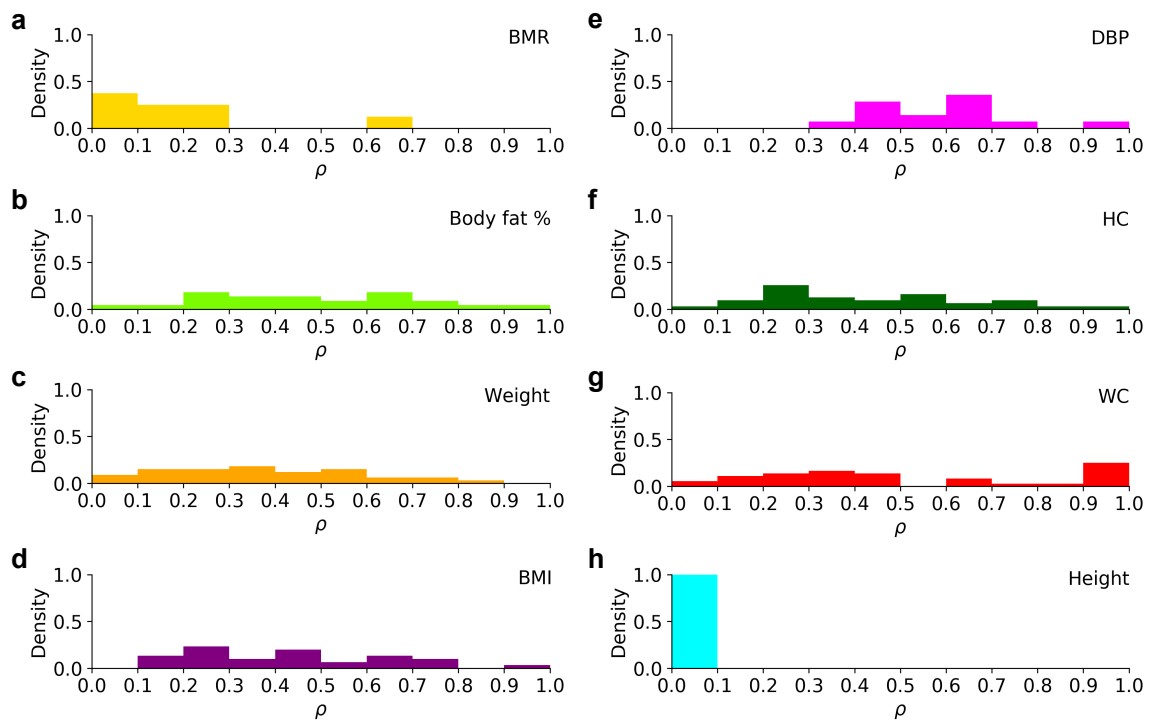
A second interesting locus lies in the *AS3MT* gene which is significantly associated with height ( $P = 7.06 \times 10^{-10}$ ), SBP ( $P = 1.20 \times 10^{-14}$ ) and BMR ( $P = 4.44 \times 10^{-9}$ ) and found to have a significant interaction effect for DBP ( $P$  cFDR-adjusted = 0.0426; *rs72841270* is the lead interaction variant; see Table B1, Appendix B, locus number 41). Interestingly, two studies have found strong significant associations for SBP (the variant reported by Simino *et al.* lies only 7,757 bp from the lead interaction variant found in this analysis), whilst no significant associations were reported for DBP<sup>410,411</sup>. These previous studies are consistent with the results for this analysis using the UK Biobank data for which a strong association with SBP is observed ( $P = 1.20 \times 10^{-14}$ ), whilst the corresponding  $P$  value for association with DBP was not genome-wide significant ( $P = 3.25 \times 10^{-7}$ ). These results suggests that the same locus influences DBP and SBP, two traits that have a phenotypic correlation of 0.671 (based on 235,623 UK Biobank individuals with phenotypic measurements for both traits) but that the effect on DBP is masked by the presence of interaction effects.

A further interesting locus lies within the *DOT1L* gene which is significantly associated with HC ( $P = 1.19 \times 10^{-10}$ ), height ( $P = 1.78 \times 10^{-58}$ ), weight ( $P = 2.46 \times 10^{-10}$ ) and BMR ( $P = 3.67 \times 10^{-18}$ ), again with a significant interaction effect for DBP ( $P$  cFDR-adjusted = 0.0445; *rs12981806* is the lead interaction variant; see Table B1, Appendix B, locus number 71). Variants 21,104 bp and 26,553 bp from this lead interaction variant have been reported as associations for pulse pressure (the difference between DBP and SBP) and 94,139 bp away for myocardial infarction<sup>412–414</sup>. These previous findings together with evidence for an association with DBP ( $P = 3.90 \times 10^{-7}$ , marginally below the genome-wide significance threshold) based on UK Biobank data, suggest that this locus has an affect on DBP, which is again masked by the presence of  $G \times E$  effects.

#### 4.3.5 Distribution of $\rho$ across different traits

As well as the identification of interaction variants across multiple phenotypes, I can also explore the relative importance of the  $G \times E$  effects across the different phenotypes. This can be achieved via the estimation of  $\rho$  (defined as the fraction of the genetic variance explained by  $G \times E$  effects at a given variant; see Section 3.2.1 for method details) per significant locus-trait pair at the lead interaction variant (smallest cFDR-adjusted  $P$  value  $< 0.05$ ). This analysis reveals that for the single

locus with a significant interaction effect on height and the majority of BMR loci identified with significant  $G \times E$  effects, the variants largely have the same effect across all individuals, with only a small component that is environmentally dependent (Fig. 4.6a and h). In comparison, loci identified with significant interaction effects for other traits have a larger component of the genetic variance explained by interaction effects (Fig. 4.6). This is particularly true of DBP, where for nine of the 14 loci,  $G \times E$  effects explain more of the genetic variance than persistent genetic effects ( $\rho > 0.5$ ; Fig. 4.6e). These results indicate that the relative importance of  $G \times E$  effects compared to persistent genetic effects varies across different traits, with  $G \times E$  effects relatively unimportant for height and BMR but that they are particularly important for DBP.



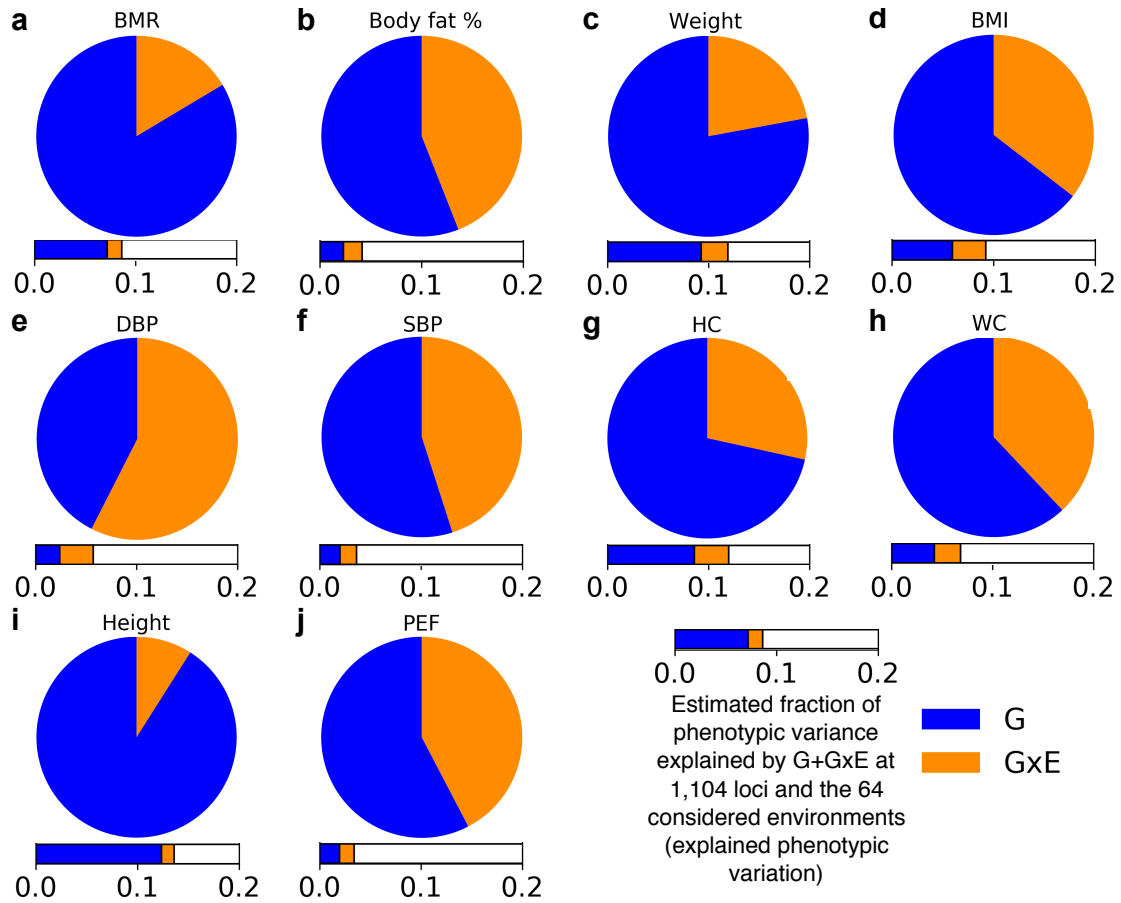
**Fig. 4.6 Distribution of  $\rho$  for the considered phenotypes** | Distribution of  $\rho$  at the lead interaction variant per locus-trait pair with an identified significant interaction effect ( $P$  cFDR-adjusted  $< 0.05$ ) for (a) BMR, (b) body fat %, (c) weight, (d) BMI, (e) DBP, (f) HC, (g) WC and (h) height (SBP and PEF are not included since there were no loci with significant  $G \times E$  effects).  $\rho = 0$  means that the effect of a variant is the same across all individuals (i. e. no  $G \times E$  effects present), whilst  $\rho = 1$  means that a variant only exerts its effect through interactions with the environment.

### 4.3.6 Relative importance of $G \times E$ effects compared to persistent genetic effects

In addition to focusing on the relative importance of  $G \times E$  effects at individual loci (see Section 4.3.5), I can also consider a more global picture by estimating the aggregate  $G \times E$  effect for the 64 considered environments across the selected 1,104 loci and compare this to the aggregate persistent genetic effect at the same set of loci (see Section 4.2.5 for method details). This approach again confirms that  $G \times E$  effects on height and BMR are small compared to persistent genetic effects (Fig. 4.7a and i);  $G \times E$  effects explain only 0.0122 (height) and 0.0142 (BMR) of the phenotypic variation whilst persistent genetic effects explain 0.1236 (height) and 0.0721 (BMR) of the phenotypic variation, such that  $G \times E$  effects explain only 0.0901 (height) and 0.1645 (BMR) of the total explained phenotypic variation (these fractions can be thought of as global estimates of  $\rho$ ). A larger proportion of the estimated explained phenotypic variation is accounted for by  $G \times E$  effects for weight (0.2209), HC (0.2843), BMI (0.3546), WC (0.3801), body fat % (0.4403) and SBP (0.4504), with  $G \times E$  effects accounting for more of the total phenotypic variation explained than persistent genetic effects for DBP at the considered 1,104 loci (Fig. 4.7). Explicitly,  $G \times E$  effects explain 0.0328 of DBP phenotypic variation whilst persistent genetic effects only explain 0.0242 of the phenotypic variation, such that  $G \times E$  effects explain 0.5751 of the total estimated phenotypic variation that is explained by both persistent genetic and  $G \times E$  effects.

With the approach taken here, it can be seen that the absolute fraction of phenotypic variation explained by persistent genetic effects is marginally lower for BMR (0.0721) than for HC (0.0858) and weight (0.0928), which could be viewed as contradicting the publicly available heritability estimates based on UK Biobank data referenced in Section 4.3.1. However, there are several plausible explanations to explain this discrepancy.

One possibility is a differing definition of phenotypic variation explained as used in this work to the publicly available heritability estimates. For the latter estimates, some covariates are regressed from the phenotype such that the phenotypic variation does not include the effect of these covariates, whilst I assume that covariates explain some phenotypic variation. The amount of phenotypic variation explained by covariates can vary across traits, which can result in some reordering of the amount of phenotypic variation explained by persistent genetic effects. The primary reason for not pre removing the effect of covariates in this analysis is that there is some subjectivity surrounding what constitutes a covariate versus an environment,



**Fig. 4.7 Estimated fraction of phenotypic variation explained by persistent genetic and G×E effects** | Bars display the estimated absolute fraction of the phenotypic variation explained by persistent genetic (G) in blue and G×E effects in orange based on lead variants across 1,104 loci and for 64 environments, with the summed total of these two absolute fractions, an estimate of the total phenotypic variation explained. The pie charts display the relative fraction of this estimated explained phenotypic variation due to persistent genetic (G) in blue and G×E effects in orange for (a) BMR, (b) body fat %, (c) weight, (d) BMI, (e) DBP, (f) SBP, (g) HC, (h) WC, (i) height and (j) PEF.

e.g. age. It should be noted that taking the effect of covariates into account has no bearing on the estimates of the fraction of the estimated explained phenotypic variation due to  $G \times E$  effects.

A second reason is that these estimates are based on a fraction of the genome (57,328 variants), whilst the previous heritability estimates are genome-wide estimates. The use of lead variants per loci (using stringent criteria to define loci) could further compound this effect as I am not accounting for any secondary effects and underestimation of the phenotypic variation explained at loci can occur if the selected lead variant is tagging the causal variant. This effect may be further exacerbated since the loci are defined based on the BMR association results. Thus these estimates provide a lower bound on the amount of phenotypic variation explained by both the persistent and  $G \times E$  components. If the amount of underestimated heritability varies substantially between the persistent genetic and  $G \times E$  components, this may effect the proportion of genetic variance explained by  $G \times E$ . As a result further work is still needed to confirm the estimates provided in this section.

## 4.4 Summary and discussion

In this chapter, I have applied the StructLMM interaction test to identify interaction effects across multiple traits at variants that are significantly associated with BMR ( $P < 5 \times 10^{-8}$ ), considering a set of 64 lifestyle based environments. This analysis was combined with a cFDR multiple testing correction, which has not been previously applied to the interaction test setting. To the best of my knowledge, this is the first analysis to explore loci with  $G \times E$  effects across different phenotypes and to estimate the amount of phenotypic variation explained by  $G \times E$  effects jointly accounting for multiple environmental variables.

I first demonstrate that application of the cFDR multiple testing correction (at a 5% threshold) to genome-wide results for BMI can dramatically increase the number of loci (from eight to 60) identified with significant interaction effects compared to using the widely employed FDR multiple testing correction (also at a 5% threshold; Fig. 4.4c and d). I note that use of the cFDR still enables the detection of strong interaction effects at variants which display little to no evidence of persistent association effects, whilst also identifying many additional variants that have weak to strong evidence for associations. This approach would also alleviate the observation in Chapter 3 that additional loci (with strong  $G \times E$  effects) to those detected

using standard LMMs were detected using the StructLMM joint association test, suggesting that existing methods that select variants for interaction testing based on association P values from previous studies or an association scan in the same dataset are not optimal.

I then test for interaction effects at 57,328 variants with 64 lifestyle based factors on ten phenotypes, including two traits, height and PEF, which act as negative controls since I do not expect to find interactions effects on these traits for the considered set of variants and environments. Appropriate multiple testing correction for phenome-wide studies is not well established, with the majority of previous studies considering both a relaxed global 10% FDR across all tested variant-trait pairs and a per variant 5% Bonferroni or FDR correction across all tested traits. Application of the cFDR to all variant-trait pairs tested, is likely to result in greater biases than that already noted in PHEWAS that apply a uniform FDR to all tested associations<sup>158</sup>. This is likely to arise due to differences in the strength of associations and enrichment of interaction effects amongst association signals across different traits. I therefore use a more stringent (than the usual 10% FDR) 5% cFDR per trait, noting that the traits included in this analysis are not independent of one another, such that a further Bonferroni correction for the number of traits would be too stringent. As this work is not predominantly variant focussed but rather an initial exploration of the interaction landscape across different phenotypes, I do not consider a per variant correction in this work. I note that application of the cFDR per variant is not valid as there are too few data points to empirically learn the dependence of interaction effects on association effects at each variant.

The interaction analysis for the ten considered traits reveals 74 unique interaction loci (two phenotypes, SBP and PEF have no significant interaction loci and height has only one significant interaction locus). 32.43% of these identified interaction loci are shared across at least three phenotypes although this set of shared phenotypes does vary at different loci, suggesting that the presence of interaction hotspots that do not solely occur due to trait correlation. Additionally, there are differences in the relative importance of  $G \times E$  effects compared to persistent genetic effects across the different traits, both when examining individual loci and when examining the 1,104 loci in aggregation.

In particular, there exist multiple lines of evidence, suggesting that  $G \times E$  effects are particularly important compared to persistent genetic effects for DBP. Explicitly, 42.86% of the identified DBP loci with significant interaction effects are not significant associations for this trait. Furthermore, the minimum estimated value

of  $\rho$  across the significant identified loci with interaction effects is 0.361, with  $G \times E$  effects explaining more of the genetic variance than persistent genetic effects ( $\rho > 0.5$ ) at nine of the 14 identified loci. Similarly, the aggregate  $G \times E$  effect across the 1,104 loci examined in this analysis, based on the lead variants, explain more phenotypic variation than persistent genetic effects. Furthermore, early GWAS, including the WTCCC, conducted on diastolic blood pressure and hypertension were underpowered to find significantly associated variants, compared to other traits<sup>57,415–418</sup>. Together these results suggest that many variants that affect DBP are masked during association analyses that assume the effect of a variant is identical across all individuals, due the presence of moderate to strong  $G \times E$  effects. As a result, in order to identify the variants that influence DBP, either very large sample sizes are required for persistent effect association analyses or  $G \times E$  effects need to be accounted for. Interestingly, there is evidence from previous studies examining different populations to suggest that DBP and BMR are correlated independently of body size<sup>419–421</sup>. In UK Biobank, the phenotypic correlation between BMR and DBP based on 330,198 individuals with measurements available for both traits is 0.2313; yet only 4.26% of loci that are significantly associated ( $P < 5 \times 10^{-8}$ ) with BMR are significant associations ( $P < 5 \times 10^{-8}$ ) for DBP (Fig. 4.3). This suggests that there may be shared environmental risk factors that have quite a large impact on both traits and/or that genetic variants that impact one of the two traits impact the second trait once environmental exposures are accounted for. The analysis conducted in this chapter does provide some evidence for the latter; the difference in the amount of phenotypic variation explained by persistent genetic effects for BMR and DBP based on the 1,104 loci that are associated ( $P < 5 \times 10^{-8}$ ) with BMR is 4.79%, with this difference in explained phenotypic variation reducing to 2.93% when  $G \times E$  effects at the 1,104 loci are also accounted for.

Whilst, it is exciting to observe traits for which  $G \times E$  effects are more important than persistent genetic effects, it is important to note that these are preliminary results that warrant further investigation. In particular, these results are based on the lead variant at each locus and therefore provide lower bound estimates on the amount of phenotypic variation explained by persistent and  $G \times E$  effects. These estimates are likely to increase if all variants at these loci are considered due to secondary effects and improved tagging of the causal variant. It will also be interesting to see if the patterns observed in this analysis hold when all variants genome-wide are considered.

Methods to estimate the phenotypic variation explained based on all variants that account for LD between the variants do exist, namely GCTA, LDAK and

LDSC<sup>68,303,422</sup>. Due to the large sample size of UK Biobank, LDSC<sup>68</sup> which utilises summary statistics would be the most appropriate. Whilst currently designed to estimate the phenotypic variation explained by persistent genetic effects, the implemented partitioned LDSC<sup>423</sup> that estimates the amount of phenotypic variation explained by different functional SNP classes could be readily adapted such that one partition accounts for the persistent genetic effects and the other for the  $G \times E$  effects. However, a major caveat of using this method is that it relies on a precomputed LD reference panel; correlation between variants multiplied by the environmental exposure can be quite different to correlation between the variants. One potential possibility is to precompute a correlation reference panel under the prior assumption that all environments are equally likely to interact with the genetic variants, the same prior assumption made in the interaction test i.e.  $\text{Corr}(\mathbf{x}_i \times \sum_{l=1}^L \mathbf{e}_l, \mathbf{x}_j \times \sum_{l=1}^L \mathbf{e}_l)$ , where  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are dosage vectors of two genetic variants and  $\mathbf{e}_l$  is the  $l^{\text{th}}$  environmental exposure vector. I note that this method can be less accurate than methods that use raw data, in particular when the number of SNPs on which the estimates are based is small, with a recommendation that at least 600,000 SNPs are used (this recommendation can be found at <https://github.com/bulik/ldsc/wiki/FAQ>).

Therefore, whilst this LDSC based approach may work if applied to SNPs genome-wide, it is unlikely to be appropriate for the 57,318 variants considered here. To overcome this I could consider using LDAK, which also has functionalities to include multiple partitions, although I will likely need to consider a small subset of individuals (in the range of 10,000 – 30,000 individuals) for computational reasons. This relatively small sample size may not be sufficient to obtain stable estimates of the phenotypic variance explained by  $G \times E$ <sup>424</sup>. Additionally, to estimate the amount of phenotypic variation explained by this approach, it may be necessary to consider further partitions of the variants with different variance components, due to possible differences in the architecture of the  $G \times E$  landscape compared to the persistent genetic landscape. Incorporation of the adaptive MultiBLUP method, which automatically identifies classes of SNPs with different effect sizes, may be appropriate to achieve this<sup>425</sup>.

A more challenging problem that would be interesting to explore is the correlation between traits in the  $G \times E$  domain analogous to existing genetic correlation approaches, or alternatively genetic correlation when accounting for both persistent and  $G \times E$  effects<sup>426</sup>. This calculation with existing methods such as LDSC requires a signed summary statistic per variant which corresponds to the persistent SNP effect direction (or alternatively can also be thought of as the average direction of the



variant effect across all individuals). However, for  $G \times E$  analysis using StructLMM we have a variant effect per individual such that the correlation calculation is now a 3D problem rather than a 2D problem, as is the case when considering persistent genetic effects.

Further avenues of exploration include increasing the set of traits analysed; in particular, inclusion of type 2 diabetes would be an interesting addition. Finally, in many cases, I observe that the lead association and lead interaction variants are not at the same chromosomal position and that the genetic variance due to persistent effects attenuates at a slower rate than the genetic variance explained by  $G \times E$  effects. It would be interesting to explore why this might be the case. Two possible hypotheses that spring to mind are (i) the correlation across variants multiplied by environmental exposures have a shorter range than the correlation based only on variant dosages, perhaps enabling  $G \times E$  tests to act as a fine mapping tool and/or (ii) that identified persistent genetic effects are tagging multiple  $G \times E$  effects. Alternatively, this may be due to artefacts such as the strength of the association or interaction effect.

Together, these results provide a first insight of the  $G \times E$  landscape across different phenotypes and demonstrate that  $G \times E$  effects can have a relatively large impact on trait outcomes, highlighting the potential importance of studying such interaction effects.



# Chapter 5

## Concluding remarks

This thesis focusses on method development and application, to test for  $G \times E$  effects that may be driven by multiple environments. Despite knowledge of interaction effects existing for more than a century<sup>427</sup> and the commonly held belief that interaction effects will explain additional phenotypic variation<sup>88,389–391</sup>, far fewer  $G \times E$  effects have been identified than persistent genetic effects. As discussed in Section 1.3.2,  $G \times E$  studies are underpowered compared to association analyses of identical sample size, predominantly due to inherent difficulties in ascertaining environmental data compared to variant information. This includes a lack of well-defined standards and an unbounded ‘global’ environmental domain (both in space and time).

It is likely that these shortcomings and the subsequent perceived ‘lack of success’ has fuelled less interest in conducting interaction studies compared to association studies, despite their potential utility in improving understanding of the mechanisms underlying complex traits and diseases. This may partially explain the low number of reported significant interaction effects.

Whilst there is further progress to be made, there have been recent efforts to improve the quality of the environmental information collected for large cohorts of individuals. This is not limited to UK Biobank which was used throughout this thesis; large population based biobanks exist in other countries, for example, deCODE in Iceland (<https://www.decode.com>), Kaiser Permanente Research Bank in the USA (<https://researchbank.kaiserpermanente.org>) and the Estonian Biobank (<https://www.geenivaramu.ee/en/access-biobank>). Therefore methods that deal and take advantage of this wealth of data are now required.

In Chapter 2, I describe a new method, StructLMM, that tests for interaction

effects at multiple environments or association effects in the presence of multiple environmental exposures. I demonstrate that unlike other multi-environment interaction and association tests, StructLMM is calibrated across a wide range of settings. In addition, I exhibit the power gains of this method over existing tests that consider a single environmental variable. This is most extreme when multiple environments contribute to the interaction effect and  $G \times E$  effects at a given variant are moderate to strong. Further advantages of this method compared to testing for effects at multiple environments one by one, are a reduced multiple testing burden and improved interpretation of the identified interactions. In particular, as highlighted in Chapter 3, the multiple environments used for interaction testing may be acting as an improved proxy for the causal environmental variable (which may be a composite of environmental factors and/or unmeasured factors) driving the observed interaction effects. Therefore, exploration of the individual environments driving an observed interaction effect is perhaps less pertinent than the identification of environmental states and thus individuals within the population that are at increased trait risk if they carry the risk increasing alleles. A future methodological improvement specific to StructLMM is the integration of methods that impute missing environmental data to increase sample size and thus power, particularly advantageous when many environmental variables are included in the analysis (as discussed in Section 3.5). Furthermore, the method could be extended and generalised, such that it can be applied to binary traits (see Section 2.5).

Initial application of StructLMM in Chapter 3, accounting for 64 lifestyle based environments, revealed a number of novel interaction and association effects on BMI using the UK Biobank data. One interesting result that perhaps warrants further exploration is the possible opposite direction of effect observed at *PARK2*. It was initially found to be associated with BMI by the GIANT consortium and later linked to BMI change during a 10 year follow up study<sup>316,331</sup>. Little is known about the precise functional properties of the *PARK2* gene, which encodes the Ubiquitin E3 ligase Parkin<sup>428</sup>. However there is a growing body of evidence that *PARK2* plays a critical role in mitochondrial homeostasis<sup>429–434</sup>. Of greater interest, a *PARK2* knock out mouse study found no difference in body weight or adiposity between wild type and knockout mice on a regular chow-fed diet but observed that knockout mice were resistant to weight gain on a high fat diet<sup>434,435</sup>; it is thought that this is due to reduced intestinal lipid absorption in *PARK2* knock out mice<sup>434,435</sup>. Exploration of the environments driving this observed interaction effect using Bayes factors in Chapter 3, suggests that tea intake is the most relevant environmental factor. Upon closer examination of the association landscape for

tea intake using publicly available data (<https://biobankengine.stanford.edu/coding/INI1488>)<sup>399</sup>, I note a number of variants are strongly associated with this environment. It is therefore plausible that the observed opposite direction of effect is capturing an instance of epistasis. Identification of epistatic effects may be functionally informative; for example it may increase knowledge of how *PARK2* gene expression is regulated or which other gene products (proteins) the *PARK2* gene product interacts with. Alternatively, if epistatic interactions can be ruled out then the observed environmentally dependent opposite direction of effect at *PARK2* (see Fig. 3.14) may be of clinical importance; specifically lifestyle changes based on the genotype that an individual carries may have a relatively large impact on BMI risk.

A second interesting result from Chapter 3, is that across the vast majority of identified interaction loci, the same individuals are not at the extremes of the trait risk spectrum. This suggests, that the environmental exposures driving the interaction effects differ across the loci and subsequently, lifestyle modifications with the greatest impact should be tailored to the individual based on the genetic risk variants that they carry, somewhat a realisation of personalised medicine. That is not to say that all individuals within a population will not benefit from following a healthy lifestyle. This also implies that existing  $G \times E$  burden based approaches (see Section 2.2.7), including the use of GRS are likely underpowered and in general not appropriate.

I demonstrate in Chapter 4 that the StructLMM interaction test can be combined with a cFDR multiple testing correction, boosting the number of identified interaction effects for a given expected false discovery rate. Specifically, the number of loci ( $\pm 1$  Mb, LD  $r^2 < 0.1$ ) identified with significant  $G \times E$  effects on BMI when testing 7,515,856 SNPs genome-wide (using the StructLMM interaction test) increased from 8 to 60 using 5% FDR (Benjamini-Hochberg) and 5% conditional FDR multiple testing adjustments, respectively. Whilst originally developed for application to two independent association scans, its application in this setting, conditioning on association results is novel. I note that this method is not only applicable to use in combination with StructLMM but can be applied to existing interaction tests that consider a single environmental variable. This method can be viewed as a soft thresholding version of existing methods that select variants to take forward for interaction testing based on marginal association results using a defined threshold. An alternative, that was not considered in this work but could be a future direction for exploration is to consider a cFDR approach that conditions on variance heterogeneity results, analogous to current variance heterogeneity hard thresholding

approaches e.g. the recently developed method by Young *et al.*<sup>436</sup>. However, unless conditioning on variance heterogeneity rather than marginal genetic effects, yields substantial increases in power, the approach currently considered has greater practicality since association scans are usually performed as a default first analysis and no further work is required to use these results as covariates.

Application of the StructLMM interaction test to a broader range of phenotypes, in Chapter 4, enabled initial exploration of the gene-environment interaction landscape across different traits. Whilst it was not necessarily surprising to see that  $G \times E$  effects explain additional phenotypic variance, comparison with the phenotypic variance explained by persistent genetic effects highlights the potential importance of interaction effects, such that they deserve greater attention in the future.

The work described in Chapter 4 is preliminary, with further work required to confirm the reported findings. In particular, adaptation and development of existing methods such as LDAK<sup>424</sup> is required to take into account the effect of all variants at the 1,104 considered loci to confirm or refute the amount of phenotypic variation explained by persistent and  $G \times E$  effects. Having said this, estimation of the fraction of genetic variance explained by  $G \times E$  effects ( $\rho$ ) at individual significant interaction loci does suggest that these results are likely to hold true. It would also be interesting to examine if the observations in Chapter 4 based on the BMR associated variants ( $P < 5 \times 10^{-8}$ ) are consistent when considering all variants genome-wide.

Future directions related to this phenome-wide interaction analysis are to explore  $G \times E$  overlap, similar to definitions of genetic correlation<sup>426</sup>. In addition, further exploration of why lead persistent and interaction genetic variants at a locus differ would be interesting; explicitly, determining whether the observed effects are tagging the same signal or if association effects are tagging multiple interaction signals would be of value to the field.

Whilst throughout this thesis, I only apply these methods to the analysis of  $G \times E$  effects, these approaches could equally be applied to explore epistasis or other burgeoning areas, such as viral-host or microbiota-host genetics, both of which have the potential to yield insightful and interesting findings.

The work in this thesis addresses some of the existing problems related to interaction tests, namely dealing with the depth of environmental data available and the use of multiple testing methods that increase the number of identified interaction loci at a given FDR. However, there are still problems that exist that are not specific to StructLMM and/or the cFDR multiple testing correction methods.

The first is the scope of the environmental data collected. In particular global standards and definitions need to be agreed and environmental data needs to be collected more densely in time and over lifespans. An advantage of continual environmental data collection is that behavioural changes due to perceived ‘monitoring’ will not be an issue, reducing the noise of interaction tests. However, even if these problems are overcome, the collection of this data will take decades.

A further, more feasible avenue of interaction method development is improved specification of the null model. Specifically, this would encompass inclusion of non-linear additive environmental terms within the null model, perhaps through the use of machine learning methods such as the random forest approach<sup>437</sup>.

Overall, the work in this thesis demonstrates the necessity and advantages of jointly modelling  $G \times E$  effects at multiple environments, providing a robust computationally efficient novel method StructLMM to do so. Such methods combined with the increasing availability of large scale biobanks have the potential to further advance our understanding of complex traits and diseases.





# References

- [1] Brownlee, J. The Inheritance of Complex Growth Forms, such as Stature, on Mendel's Theory. *Int. J. Epidemiol.* **42**, 932–934 (2013).
- [2] Fisher, R. A. XV.-The Correlation between Relatives on the Supposition of Mendelian Inheritance (1919). *Trans. R. Soc. Edinburgh* **52**, 399–433 (2012).
- [3] Pearson, K. & Lee, A. On the Laws of Inheritance in Man: I. Inheritance of Physical Characters. *Biometrika* **2**, 357 (1903).
- [4] Olby, R. The Dimensions of Scientific Controversy: The Biometric-Mendelian Debate. *Br. J. Hist. Sci.* **22**, 299 (1989).
- [5] Visscher, P. M. Commentary: Height and Mendel's theory: the long and the short of it. *Int. J. Epidemiol.* **42**, 944–945 (2013).
- [6] Provine, W. B. The origins of theoretical population genetics. 211 (University of Chicago Press, 2001).
- [7] Spencer, H. G. & Paul, D. B. The failure of a scientific critique: David Heron, Karl Pearson and Mendelian eugenics. *Br. J. Hist. Sci.* **31**, 441–52 (1998).
- [8] Stranger, B. E., Stahl, E. A. & Raj, T. Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics. *Genetics* **187**, 367–383 (2011).
- [9] Mendel, G. Experiments in plant hybridisation (1865). Tech. Rep. (1996).
- [10] Galton, F. Hereditary genius: An inquiry into its laws and consequences. (Cosimo Classics, 1869).
- [11] Stratthdee, C. A., Gavish, H., Shannon, W. R. & Buchwald, M. Cloning of cDNAs for Fanconi's anaemia by functional complementation. *Nature* **356**, 763–767 (1992).

- [12] Riordan, J. R. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245**, 1066–73 (1989).
- [13] Kerem, B. *et al.* Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073–80 (1989).
- [14] Koenig, M. *et al.* Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. *Cell* **50**, 509–17 (1987).
- [15] De Paepe, A., Nuytinck, L., Hausser, I., Anton-Lamprecht, I. & Naeyaert, J. M. Mutations in the COL5A1 gene are causal in the Ehlers-Danlos syndromes I and II. *Am. J. Hum. Genet.* **60**, 547–54 (1997).
- [16] Wenstrup, R. J., Langland, G. T., Willing, M. C., D’Souza, V. N. & Cole, W. G. A splice-junction mutation in the region of COL5A1 that codes for the carboxyl propeptide of pro alpha 1(V) chains results in the gravis form of the Ehlers-Danlos syndrome (type I). *Hum. Mol. Genet.* **5**, 1733–6 (1996).
- [17] Badano, J. L. & Katsanis, N. Beyond Mendel: an evolving view of human genetic disease transmission. *Nat. Rev. Genet.* **3**, 779–789 (2002).
- [18] Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33**, 228–237 (2003).
- [19] Scriver, C. R. The PAH gene, phenylketonuria, and a paradigm shift. *Hum. Mutat.* **28**, 831–845 (2007).
- [20] Ottman, R. Gene-environment interaction: definitions and study designs. *Prev. Med. (Baltim).* **25**, 764–70 (1996).
- [21] Spataro, N., Rodríguez, J. A., Navarro, A. & Bosch, E. Properties of human disease genes and the role of genes linked to Mendelian disorders in complex disease aetiology. *Hum. Mol. Genet.* **26**, 489–500 (2017).
- [22] Price, A. L., Spencer, C. C. A. & Donnelly, P. Progress and promise in understanding the genetic basis of common diseases. *Proceedings. Biol. Sci.* **282**, 20151684 (2015).
- [23] Mackay, T. F. C. Q&A: Genetic analysis of quantitative traits. *J. Biol.* **8**, 23 (2009).

- [24] Rowe, S. J. & Tenesa, A. Human complex trait genetics: lifting the lid of the genomics toolbox - from pathways to prediction. *Curr. Genomics* **13**, 213–24 (2012).
- [25] Burton, P. R., Tobin, M. D. & Hopper, J. L. Key concepts in genetic epidemiology. *Lancet* **366**, 941–951 (2005).
- [26] Bush, W. S. & Moore, J. H. Chapter 11: Genome-Wide Association Studies. *PLoS Comput. Biol.* **8**, e1002822 (2012).
- [27] van Heyningen, V. & Yeyati, P. L. Mechanisms of non-Mendelian inheritance in genetic disease. *Hum. Mol. Genet.* **13**, R225–R233 (2004).
- [28] Dipple, K. M. & McCabe, E. R. Phenotypes of patients with "simple" Mendelian disorders are complex traits: thresholds, modifiers, and systems dynamics. *Am. J. Hum. Genet.* **66**, 1729–35 (2000).
- [29] Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–31 (1980).
- [30] Pulst, S. M. Genetic Linkage Analysis. *Arch. Neurol.* **56**, 667 (1999).
- [31] Dawn Teare, M. & Barrett, J. H. Genetic linkage studies. *Lancet* **366**, 1036–1044 (2005).
- [32] Lander, E. & Kruglyak, L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat. Genet.* **11**, 241–247 (1995).
- [33] Rahman, N. *et al.* Ehlers-Danlos syndrome with severe early-onset periodontal disease (EDS-VIII) is a distinct, heterogeneous disorder with one predisposition gene at chromosome 12p13. *Am. J. Hum. Genet.* **73**, 198–204 (2003).
- [34] Hampe, J. *et al.* Association between insertion mutation in NOD2 gene and Crohn's disease in German and British populations. *Lancet* **357**, 1925–1928 (2001).
- [35] Hugot, J.-P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
- [36] Ogura, Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).

- [37] Brant, S. R. *et al.* American families with Crohn’s disease have strong evidence for linkage to chromosome 16 but not chromosome 12. *Gastroenterology* **115**, 1056–61 (1998).
- [38] Bailey-Wilson, J. E. & Wilson, A. F. Linkage analysis in the next-generation sequencing era. *Hum. Hered.* **72**, 228–36 (2011).
- [39] Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**, 66–71 (1994).
- [40] Nyholt, D. R. All LODs are not created equal. *Am. J. Hum. Genet.* **67**, 282–8 (2000).
- [41] Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLoS Med.* **2**, e124 (2005).
- [42] Grond-Ginsbach, C. *et al.* Exclusion mapping of the genetic predisposition for cervical artery dissections by linkage analysis. *Ann. Neurol.* **52**, 359–364 (2002).
- [43] Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–10 (2001).
- [44] Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
- [45] Bush, W. S. & Haines, J. Overview of Linkage Analysis in Complex Traits. In *Curr. Protoc. Hum. Genet.*, vol. 64, 1.9.1–1.9.18 (John Wiley & Sons, Inc., 2010).
- [46] Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–7 (1996).
- [47] Jorde, L. B. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10**, 1435–44 (2000).
- [48] Cordell, H. J. & Clayton, D. G. Genetic association studies. *Lancet* **366**, 1121–1131 (2005).
- [49] Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
- [50] Gibbs, R. A. *et al.* The International HapMap Project. *Nature* **426**, 789–796 (2003).

- [51] International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
- [52] Klein, R. J. *et al.* Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science* **308**, 385–389 (2005).
- [53] Yamazaki, K. *et al.* Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn’s disease. *Hum. Mol. Genet.* **14**, 3499–3506 (2005).
- [54] Ozaki, K. & Tanaka, T. Genome-wide association study to identify SNPs conferring risk of myocardial infarction and their functional analyses. *Cell. Mol. Life Sci.* **62**, 1804–1813 (2005).
- [55] Duerr, R. H. *et al.* A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. *Science* **314**, 1461–1463 (2006).
- [56] Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
- [57] Wellcome Trust Case Control Consortium, T. W. T. C. C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–78 (2007).
- [58] Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
- [59] Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18**, 77 (2017).
- [60] Evangelou, E. & Ioannidis, J. P. A. Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* **14**, 379–389 (2013).
- [61] Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet.* **40**, 189–197 (2008).
- [62] Lettre, G., Lange, C. & Hirschhorn, J. N. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet. Epidemiol.* **31**, 358–362 (2007).
- [63] Rao, C. R. Score Test: Historical Review and Recent Developments. In *Adv. Rank. Sel. Mult. Comp. Reliab.*, 3–20 (Birkhäuser Boston, Boston, MA, 2005).

- [64] Xing, G., Lin, C., Wooding, S. P. & Xing, C. Blindly Using Wald’s Test Can Miss Rare Disease-Causal Variants in Case-Control Association Studies. *Ann. Hum. Genet.* **76**, 168–177 (2012).
- [65] Palmer, L. J. & Cardon, L. R. Shaking the tree: mapping complex disease genes with linkage disequilibrium. *Lancet* **366**, 1223–1234 (2005).
- [66] Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
- [67] Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
- [68] Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
- [69] McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).
- [70] Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
- [71] Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. Association Mapping in Structured Populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
- [72] Price, A. L., Zaitlen, N. A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* **11**, 459–463 (2010).
- [73] Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* **42**, 348–354 (2010).
- [74] Hayes, B. J., Visscher, P. M. & Goddard, M. E. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet. Res. (Camb)*. **91**, 47 (2009).
- [75] Hoffman, G. E. Correcting for Population Structure and Kinship Using the Linear Mixed Model: Theory and Extensions. *PLoS One* **8**, e75707 (2013).
- [76] Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–5 (2011).
- [77] Loh, P. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

- [78] Lippert, C., Casale, F. P., Rakitsch, B. & Stegle, O. LIMIX: genetic analysis of multiple traits. *bioRxiv* 003905 (2014).
- [79] Casale, F. P., Rakitsch, B., Lippert, C. & Stegle, O. Efficient set tests for the genetic analysis of correlated traits. *Nat. Methods* **12**, 755–8 (2015).
- [80] Casale, F. P., Horta, D., Rakitsch, B. & Stegle, O. Joint genetic analysis using variant sets reveals polygenic gene-context interactions. *PLOS Genet.* **13**, e1006693 (2017).
- [81] Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet. Epidemiol.* **32**, 227–34 (2008).
- [82] Davis, J. *et al.* An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage Disequilibrium between Variants. *Am. J. Hum. Genet.* **98**, 216–224 (2016).
- [83] Gao, X., Starmer, J. & Martin, E. R. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet. Epidemiol.* **32**, 361–369 (2008).
- [84] Pe’er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32**, 381–385 (2008).
- [85] Zablocki, R. W. *et al.* Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics* **30**, 2098–2104 (2014).
- [86] Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
- [87] Efron, B. & Tibshirani, R. Empirical bayes methods and false discovery rates for microarrays. *Genet. Epidemiol.* **23**, 70–86 (2002).
- [88] Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–53 (2009).
- [89] Saint Pierre, A. & Genin, E. How important are rare variants in common disease? *Brief. Funct. Genomics* **13**, 353–361 (2014).
- [90] Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).

- [91] Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn’s disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).
- [92] Auer, P. L. & Lettre, G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med.* **7**, 16 (2015).
- [93] Dolled-Filhart, M. P., Lee, M., Ou-yang, C., Haraksingh, R. R. & Lin, J. C. Computational and Bioinformatics Frameworks for Next-Generation Whole Exome and Genome Sequencing. *Sci. World J.* **2013**, 1–10 (2013).
- [94] The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
- [95] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
- [96] Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
- [97] Sazonovs, A. & Barrett, J. C. Rare-Variant Studies to Complement Genome-Wide Association Studies. *Annu. Rev. Genom. Hum. Genet* **19**, 97–112 (2018).
- [98] Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
- [99] Helgadóttir, A. *et al.* Genome-wide analysis yields new loci associating with aortic valve stenosis. *Nat. Commun.* **9**, 987 (2018).
- [100] Francioli, L. C. *et al.* Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat. Genet.* **46**, 818–825 (2014).
- [101] Gilly, A. *et al.* Very low-depth sequencing in a founder population identifies a cardioprotective APOC3 signal missed by genome-wide imputation. *Hum. Mol. Genet.* **25**, 2360–2365 (2016).
- [102] Sidore, C. *et al.* Genome sequencing elucidates Sardinian genetic architecture and augments association analyses for lipid and blood inflammatory markers. *Nat. Genet.* **47**, 1272–1281 (2015).



- [103] Pistis, G. *et al.* Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *Eur. J. Hum. Genet.* **23**, 975–983 (2015).
- [104] The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
- [105] Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* **517**, 327–332 (2015).
- [106] Asimit, J. & Zeggini, E. Rare Variant Association Analysis Methods for Complex Traits. *Annu. Rev. Genet.* **44**, 293–308 (2010).
- [107] Madsen, B. E. & Browning, S. R. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genet.* **5**, e1000384 (2009).
- [108] Morris, A. P. & Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* **34**, 188–193 (2010).
- [109] Li, B. & Leal, S. M. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
- [110] Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutat. Res. Mol. Mech. Mutagen.* **615**, 28–56 (2007).
- [111] Asimit, J. L., Day-Williams, A. G., Morris, A. P. & Zeggini, E. ARIEL and AMELIA: testing for an accumulation of rare variants using next-generation sequencing data. *Hum. Hered.* **73**, 84–94 (2012).
- [112] Han, F. & Pan, W. A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants. *Hum. Hered.* **70**, 42–54 (2010).
- [113] Ionita-Laza, I., Buxbaum, J. D., Laird, N. M. & Lange, C. A New Testing Strategy to Identify Rare Variants with Either Risk or Protective Effect on Disease. *PLoS Genet.* **7**, e1001289 (2011).
- [114] Liu, D. J. & Leal, S. M. A Novel Adaptive Method for the Analysis of Next-Generation Sequencing Data to Detect Complex Trait Associations with Rare Variants Due to Gene Main Effects and Interactions. *PLoS Genet.* **6**, e1001156 (2010).

- [115] Price, A. L. *et al.* Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *Am. J. Hum. Genet.* **86**, 832–838 (2010).
- [116] Lin, D. & Tang, Z. A General Framework for Detecting Disease Associations with Rare Variants in Sequencing Studies. *Am. J. Hum. Genet.* **89**, 354–367 (2011).
- [117] Hoffmann, T. J., Marini, N. J. & Witte, J. S. Comprehensive Approach to Analyzing Rare Genetic Variants. *PLoS One* **5**, e13584 (2010).
- [118] Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
- [119] Neale, B. M. *et al.* Testing for an Unusual Distribution of Rare Variants. *PLoS Genet.* **7**, e1001322 (2011).
- [120] Pan, W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet. Epidemiol.* **33**, 497–507 (2009).
- [121] Sun, J., Zheng, Y. & Hsu, L. A Unified Mixed-Effects Model for Rare-Variant Association in Sequencing Studies. *Genet. Epidemiol.* **37**, 334–344 (2013).
- [122] Derkach, A., Lawless, J. F. & Sun, L. Robust and Powerful Tests for Rare Variants Using Fisher’s Method to Combine Evidence of Association From Two or More Complementary Tests. *Genet. Epidemiol.* **37**, 110–121 (2013).
- [123] Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–75 (2012).
- [124] Liu, D. J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* **46**, 200–204 (2014).
- [125] Cruchaga, C. *et al.* Rare coding variants in the phospholipase D3 gene confer risk for Alzheimer’s disease. *Nature* **505**, 550–554 (2014).
- [126] Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
- [127] Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–4 (2003).
- [128] Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).

- [129] Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
- [130] De Souza, Y. G. & Greenspan, J. S. Biobanking past, present and future: responsibilities and benefits. *AIDS* **27**, 303–12 (2013).
- [131] Collins, R. What makes UK Biobank special? *Lancet* **379**, 1173–1174 (2012).
- [132] Bookman, E. B. *et al.* Gene-environment interplay in common complex diseases: forging an integrative model - recommendations from an NIH workshop. *Genet. Epidemiol.* **35**, 217–25 (2011).
- [133] Korte, A. *et al.* A mixed-model approach for genome-wide association studies of correlated traits in structured populations. *Nat. Genet.* **44**, 1066–1071 (2012).
- [134] Joo, J. W. J. *et al.* Efficient and Accurate Multiple-Phenotype Regression Method for High Dimensional Data Considering Population Structure. *Genetics* **204**, 1379–1390 (2016).
- [135] Furlotte, N. A. & Eskin, E. Efficient Multiple-Trait Association and Estimation of Genetic Correlation Using the Matrix-Variate Linear Mixed Model. *Genetics* **200**, 59–68 (2015).
- [136] Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **11**, 407–409 (2014).
- [137] Stearns, F. W. & Cairns, J. One hundred years of pleiotropy: a retrospective. *Genetics* **186**, 767–73 (2010).
- [138] Hackinger, S. & Zeggini, E. Statistical methods to detect pleiotropy in human complex traits. *Open Biol.* **7** (2017).
- [139] Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89**, 607–18 (2011).
- [140] Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).
- [141] Gage, S. H., Davey Smith, G., Ware, J. J., Flint, J. & Munafò, M. R.  $G = E$ : What GWAS Can Tell Us about the Environment. *PLOS Genet.* **12**, e1005765 (2016).

- [142] Bennett, D. A. & Holmes, M. V. Mendelian randomisation in cardiovascular research: an introduction for clinicians. *Heart* **103**, 1400–1407 (2017).
- [143] Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–98 (2014).
- [144] Zhernakova, A., van Diemen, C. C. & Wijmenga, C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat. Rev. Genet.* **10**, 43–55 (2009).
- [145] Criswell, L. A. *et al.* Analysis of Families in the Multiple Autoimmune Disease Genetics Consortium (MADGC) Collection: the PTPN22 620W Allele Associates with Multiple Autoimmune Phenotypes. *Am. J. Hum. Genet.* **76**, 561–571 (2005).
- [146] Ueda, H. *et al.* Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* **423**, 506–511 (2003).
- [147] Helgadottir, A. *et al.* A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction. *Science* **316**, 1491–1493 (2007).
- [148] Samani, N. J. *et al.* Genomewide Association Analysis of Coronary Artery Disease. *N. Engl. J. Med.* **357**, 443–453 (2007).
- [149] McPherson, R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488–91 (2007).
- [150] Saxena, R. *et al.* Genome-Wide Association Analysis Identifies Loci for Type 2 Diabetes and Triglyceride Levels. *Science* **316**, 1331–1336 (2007).
- [151] Bush, W. S., Oetjens, M. T. & Crawford, D. C. Unravelling the human genome-phenome relationship using phenome-wide association studies. *Nat. Rev. Genet.* **17**, 129–145 (2016).
- [152] MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
- [153] Matise, T. C. *et al.* The Next PAGE in Understanding Complex Traits: Design for the Analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study. *Am. J. Epidemiol.* **174**, 849–859 (2011).

- [154] Denny, J. C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).
- [155] Denny, J. C., Bastarache, L. & Roden, D. M. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annu. Rev. Genom. Hum. Genet* **17**, 353–73 (2016).
- [156] Cronin, R. M. *et al.* Phenome-wide association studies demonstrating pleiotropy of genetic variants within FTO with and without adjustment for body mass index. *Front. Genet.* **5**, 250 (2014).
- [157] Pendergrass, S. A. & Ritchie, M. D. Phenome-Wide Association Studies: Leveraging Comprehensive Phenotypic and Genotypic Data for Discovery. *Curr. Genet. Med. Rep.* **3**, 92–100 (2015).
- [158] Verma, A. *et al.* PheWAS and Beyond: The Landscape of Associations with Medical Diagnoses and Clinical Measures across 38,662 Individuals from Geisinger. *Am. J. Hum. Genet.* **102**, 592–608 (2018).
- [159] Hebbring, S. J. The challenges, advantages and future of phenome-wide association studies. *Immunology* **141**, 157–165 (2014).
- [160] McGinnis, D. P., Brownstein, J. S. & Patel, C. J. Environment-Wide Association Study of Blood Pressure in the National Health and Nutrition Examination Survey (1999-2012). *Sci. Rep.* **6**, 30373 (2016).
- [161] Chen, H. *et al.* Spatial Association Between Ambient Fine Particulate Matter and Incident Hypertension. *Circulation* **129**, 562–569 (2014).
- [162] Dvorchak, J. T. *et al.* Acute Effects of Ambient Particulate Matter on Blood Pressure: Differential Effects Across Urban Communities. *Hypertension* **53**, 853–859 (2009).
- [163] Tzoulaki, I. *et al.* A Nutrient-Wide Association Study on Blood Pressure. *Circulation* **126**, 2456–2464 (2012).
- [164] Everett, C. J., Mainous, A. G., Frithsen, I. L., Player, M. S. & Matheson, E. M. Association of polychlorinated biphenyls with hypertension in the 1999-2002 National Health and Nutrition Examination Survey. *Environ. Res.* **108**, 94–97 (2008).

- [165] Telišman, S., Jurasović, J., Pizent, A. & Cvitković, P. Blood Pressure in Relation to Biomarkers of Lead, Cadmium, Copper, Zinc, and Selenium in Men without Occupational Exposure to Metals. *Environ. Res.* **87**, 57–68 (2001).
- [166] Navas-Acien, A., Guallar, E., Silbergeld, E. K. & Rothenberg, S. J. Lead Exposure and Cardiovascular Disease - A Systematic Review. *Environ. Health Perspect.* **115**, 472–482 (2006).
- [167] Lee, B. & Kim, Y. Association of blood cadmium with hypertension in the Korean general population: Analysis of the 2008-2010 Korean national health and nutrition examination survey data. *Am. J. Ind. Med.* **55**, 1060–1067 (2012).
- [168] Tellez-Plaza, M., Navas-Acien, A., Crainiceanu, C. M. & Guallar, E. Cadmium Exposure and Hypertension in the 1999-2004 National Health and Nutrition Examination Survey (NHANES). *Environ. Health Perspect.* **116**, 51–56 (2007).
- [169] Peters, J. L., Patricia Fabian, M. & Levy, J. I. Combined impact of lead, cadmium, polychlorinated biphenyls and non-chemical risk factors on blood pressure in NHANES. *Environ. Res.* **132**, 93–99 (2014).
- [170] Patel, C. J., Bhattacharya, J. & Butte, A. J. An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus. *PLoS One* **5**, e10746 (2010).
- [171] Hall, M. A. *et al.* Environment-wide association study (EWAS) for type 2 diabetes in the Marshfield Personalized Medicine Research Project Biobank. *Pac. Symp. Biocomput.* 200–11 (2014).
- [172] Cordell, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **11**, 2463–2468 (2002).
- [173] Frankel, W. N. & Schork, N. J. Who's afraid of epistasis? *Nat. Genet.* **14**, 371–373 (1996).
- [174] Andreassen, C. H. *et al.* Non-Replication of Genome-Wide Based Associations between Common Variants in INSIG2 and PFKP and Obesity in Studies of 18,014 Danes. *PLoS One* **3**, e2872 (2008).
- [175] Murcray, C. E., Lewinger, J. P. & Gauderman, W. J. Gene-Environment Interaction in Genome-Wide Association Studies. *Am. J. Epidemiol.* **169**, 219–226 (2008).

- [176] Caspi, A., Hariri, A. R., Holmes, A., Uher, R. & Moffitt, T. E. Genetic Sensitivity to the Environment: The Case of the Serotonin Transporter Gene and Its Implications for Studying Complex Diseases and Traits. *Am. J. Psychiatry* **167**, 509–527 (2010).
- [177] Thomas, D. Methods for Investigating Gene-Environment Interactions in Candidate Pathway and Genome-Wide Association Studies. *Annu. Rev. Public Health* **31**, 21–36 (2010).
- [178] García-Closas, M. *et al.* NAT2 slow acetylation, GSTM1 null genotype, and risk of bladder cancer: results from the Spanish Bladder Cancer Study and meta-analyses. *Lancet* **366**, 649–659 (2005).
- [179] Gauderman, W. J. *et al.* Update on the State of the Science for Analytical Methods for Gene-Environment Interactions. *Am. J. Epidemiol.* **186**, 762–770 (2017).
- [180] Wei, W., Hemani, G. & Haley, C. S. Detecting epistasis in human complex traits. *Nat. Rev. Genet.* **15**, 722–733 (2014).
- [181] Moore, J. H. & Williams, S. M. Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.* **85**, 309–20 (2009).
- [182] Thomas, D. Gene–environment-wide association studies: emerging approaches. *Nat. Rev. Genet.* **11**, 259–72 (2010).
- [183] McAllister, K. *et al.* Current Challenges and New Opportunities for Gene-Environment Interaction Studies of Complex Diseases. *Am. J. Epidemiol.* **186**, 753–761 (2017).
- [184] Greenland, S. Interactions in Epidemiology: Relevance, Identification, and Estimation. *Epidemiology* **20**, 14–17 (2009).
- [185] W. Bateson. Mendel’s principles of heredity (Cambridge University Press, 1909).
- [186] Phillips, P. C. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9**, 855–67 (2008).
- [187] Wang, X., Elston, R. C. & Zhu, X. The Meaning of Interaction. *Hum. Hered.* **70**, 269–277 (2010).
- [188] Cordell, H. J. Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.* **10**, 392–404 (2009).

- [189] Simon, P. H., Sylvestre, M., Tremblay, J. & Hamet, P. Key Considerations and Methods in the Study of Gene-Environment Interactions. *Am. J. Hypertens.* **29**, 891–899 (2016).
- [190] Smith, P. G. & Day, N. E. The design of case-control studies: the influence of confounding and interaction effects. *Int. J. Epidemiol.* **13**, 356–65 (1984).
- [191] Hunter, D. J. Gene-environment interactions in human diseases. *Nat. Rev. Genet.* **6**, 287–298 (2005).
- [192] Ritz, B. R. *et al.* Lessons Learned From Past Gene-Environment Interaction Successes. *Am. J. Epidemiol.* **186**, 778–786 (2017).
- [193] Brown, A. A. *et al.* Genetic interactions affecting human gene expression identified by variance association mapping. *Elife* **3**, e01381 (2014).
- [194] Fairfax, B. P. *et al.* Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. *Science* **343**, 1246949–1246949 (2014).
- [195] Young, A. I., Wauthier, F. & Donnelly, P. Multiple novel gene-by-environment interactions modify the effect of FTO variants on body mass index. *Nat. Commun.* **7**, 12724 (2016).
- [196] Bjørnland, T., Langaas, M., Grill, V. & Mostad, I. L. Assessing gene-environment interaction effects of FTO, MC4R and lifestyle factors on obesity using an extreme phenotype sampling design: Results from the HUNT study. *PLoS One* **12**, e0175071 (2017).
- [197] Kilpeläinen, T. O. *et al.* Physical Activity Attenuates the Influence of FTO Variants on Obesity Risk: A Meta-Analysis of 218,166 Adults and 19,268 Children. *PLoS Med.* **8**, e1001116 (2011).
- [198] Reddon, H. *et al.* Physical activity and genetic predisposition to obesity in a multiethnic longitudinal study. *Sci. Rep.* **6**, 18672 (2016).
- [199] Graff, M. *et al.* Genome-wide physical activity interactions in adiposity - A meta-analysis of 200,452 adults. *PLOS Genet.* **13**, e1006528 (2017).
- [200] Qi, Q. *et al.* FTO genetic variants, dietary intake and body mass index: insights from 177,330 individuals. *Hum. Mol. Genet.* **23**, 6961–6972 (2014).



- [201] Phillips, C. M. *et al.* High Dietary Saturated Fat Intake Accentuates Obesity Risk Associated with the Fat Mass and Obesity-Associated Gene in Adults. *J. Nutr.* **142**, 824–831 (2012).
- [202] Corella, D. *et al.* A High Intake of Saturated Fatty Acids Strengthens the Association between the Fat Mass and Obesity-Associated Gene and BMI. *J. Nutr.* **141**, 2219–2225 (2011).
- [203] Qi, Q. *et al.* Dietary Intake, FTO Genetic Variants, and Adiposity: A Combined Analysis of Over 16,000 Children and Adolescents. *Diabetes* **64**, 2467–2476 (2015).
- [204] Ahmad, T. *et al.* Lifestyle interaction with fat mass and obesity-associated (FTO) genotype and risk of obesity in apparently healthy U.S. women. *Diabetes Care* **34**, 675–80 (2011).
- [205] Sonestedt, E. *et al.* Fat and carbohydrate intake modify the association between genetic variation in the FTO genotype and obesity. *Am. J. Clin. Nutr.* **90**, 1418–1425 (2009).
- [206] Corella, D. *et al.* Statistical and Biological Gene-Lifestyle Interactions of MC4R and FTO with Diet and Physical Activity on Obesity: New Effects on Alcohol Consumption. *PLoS One* **7**, e52344 (2012).
- [207] Tyrrell, J. *et al.* Gene-obesogenic environment interactions in the UK Biobank study. *Int. J. Epidemiol.* **46**, dyw337 (2017).
- [208] Patel, C. J. & Ioannidis, J. P. A. Placing epidemiological results in the context of multiplicity and typical correlations of exposures. *J. Epidemiol. Community Health* **68**, 1096–1100 (2014).
- [209] Gauderman, W. J. *et al.* The Effect of Air Pollution on Lung Development from 10 to 18 Years of Age. *N. Engl. J. Med.* **351**, 1057–1067 (2004).
- [210] Smith, G. D. *et al.* Clustered Environments and Randomized Genes: A Fundamental Distinction between Conventional and Genetic Epidemiology. *PLoS Med.* **4**, e352 (2007).
- [211] Ioannidis, J. P. A., Loy, E. Y., Poulton, R. & Chia, K. S. Researching Genetic Versus Nongenetic Determinants of Disease: A Comparison and Proposed Unification. *Sci. Transl. Med.* **1**, 7ps8–7ps8 (2009).

- [212] Li, S. *et al.* Physical Activity Attenuates the Genetic Predisposition to Obesity in 20,000 Men and Women from EPIC-Norfolk Prospective Population Study. *PLoS Med.* **7**, e1000332 (2010).
- [213] Ahmad, S. *et al.* Gene  $\times$  Physical Activity Interactions in Obesity: Combined Analysis of 111,421 Individuals of European Ancestry. *PLoS Genet.* **9**, e1003607 (2013).
- [214] Qi, Q. *et al.* Fried food consumption, genetic risk, and body mass index: gene-diet interaction analysis in three US cohort studies. *BMJ* **348**, g1610 (2014).
- [215] Bycroft, C. *et al.* Genome-wide genetic data on  $\sim$ 500,000 UK Biobank participants. *bioRxiv* 166298 (2017).
- [216] Kraft, P. & Hunter, D. Integrating epidemiology and genetic association: the challenge of gene-environment interaction. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **360**, 1609–16 (2005).
- [217] Aschard, H., Zaitlen, N., Lindström, S. & Kraft, P. Variation in Predictive Ability of Common Genetic Variants by Established Strata. *Epidemiology* **26**, 51–58 (2015).
- [218] Kraft, P., Yen, Y., Stram, D. O., Morrison, J. & Gauderman, W. J. Exploiting gene-environment interaction to detect genetic associations. *Hum. Hered.* **63**, 111–9 (2007).
- [219] Sun, X. *et al.* Analysis pipeline for the epistasis search - statistical versus biological filtering. *Front. Genet.* **5**, 106 (2014).
- [220] Xie, M., Li, J. & Jiang, T. Detecting genome-wide epistases based on the clustering of relatively frequent items. *Bioinformatics* **28**, 5–12 (2012).
- [221] Turner, S. D. *et al.* Knowledge-Driven Multi-Locus Analysis Reveals Gene-Gene Interactions Influencing HDL Cholesterol Level in Two Independent EMR-Linked Biobanks. *PLoS One* **6**, e19586 (2011).
- [222] Ma, L. *et al.* Knowledge-Driven Analysis Identifies a Gene-Gene Interaction Affecting High-Density Lipoprotein Cholesterol Levels in Multi-Ethnic Populations. *PLoS Genet.* **8**, e1002714 (2012).
- [223] Ritchie, M. D. Using Biological Knowledge to Uncover the Mystery in the Search for Epistasis in Genome-Wide Association Studies. *Ann. Hum. Genet.* **75**, 172–182 (2011).

- [224] Gauderman, W. J., Zhang, P., Morrison, J. L. & Lewinger, J. P. Finding Novel Genes by Testing  $G \times E$  Interactions in a Genome-Wide Association Study. *Genet. Epidemiol.* **37**, 603–613 (2013).
- [225] Hsu, L. *et al.* Powerful Cocktail Methods for Detecting Genome-Wide Gene-Environment Interaction. *Genet. Epidemiol.* **36**, 183–194 (2012).
- [226] Kooperberg, C. & LeBlanc, M. Increasing the power of identifying gene  $\times$  gene interactions in genome-wide association studies. *Genet. Epidemiol.* **32**, 255–263 (2008).
- [227] Murcray, C. E., Lewinger, J. P., Conti, D. V., Thomas, D. C. & Gauderman, W. J. Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genet. Epidemiol.* **35**, 201–210 (2011).
- [228] Dai, J. Y. *et al.* Simultaneously testing for marginal genetic association and gene-environment interaction. *Am. J. Epidemiol.* **176**, 164–73 (2012).
- [229] Dai, J. Y., Kooperberg, C., Leblanc, M. & Prentice, R. L. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika* **99**, 929–944 (2012).
- [230] Aschard, H., Zaitlen, N., Tamimi, R. M., Lindström, S. & Kraft, P. A Nonparametric Test to Detect Quantitative Trait Loci Where the Phenotypic Distribution Differs by Genotypes. *Genet. Epidemiol.* **37**, 323–333 (2013).
- [231] Paré, G., Cook, N. R., Ridker, P. M. & Chasman, D. I. On the Use of Variance per Genotype as a Tool to Identify Quantitative Trait Interaction Effects: A Report from the Women’s Genome Health Study. *PLoS Genet.* **6**, e1000981 (2010).
- [232] Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* **37**, 413–417 (2005).
- [233] Evans, D. M., Marchini, J., Morris, A. P. & Cardon, L. R. Two-Stage Two-Locus Models in Genome-Wide Association. *PLoS Genet.* **2**, e157 (2006).
- [234] Ma, L., Clark, A. G. & Keinan, A. Gene-based testing of interactions in association studies of quantitative traits. *PLoS Genet.* **9**, e1003321 (2013).
- [235] Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U. & Wacholder, S. Powerful Multilocus Tests of Genetic Association in the Presence of Gene-Gene

- and Gene-Environment Interactions. *Am. J. Hum. Genet.* **79**, 1002–1016 (2006).
- [236] Aschard, H. A perspective on interaction effects in genetic association studies. *Genet. Epidemiol.* **40**, 678–688 (2016).
  - [237] Pollin, T. I. *et al.* Genetic Modulation of Lipid Profiles following Lifestyle Modification or Metformin Treatment: The Diabetes Prevention Program. *PLoS Genet.* **8**, e1002895 (2012).
  - [238] Langenberg, C. *et al.* Gene-Lifestyle Interaction and Type 2 Diabetes: The EPIC InterAct Case-Cohort Study. *PLoS Med.* **11**, e1001647 (2014).
  - [239] Fu, Z. *et al.* Interaction of cigarette smoking and carcinogen-metabolizing polymorphisms in the risk of colorectal polyps. *Carcinogenesis* **34**, 779–86 (2013).
  - [240] Jiao, S. *et al.* SBERIA: Set-Based Gene-Environment Interaction Test for Rare and Common Variants in Complex Diseases. *Genet. Epidemiol.* **37**, 452–464 (2013).
  - [241] Jiao, S. *et al.* Powerful Set-Based Gene-Environment Interaction Testing Framework for Complex Diseases. *Genet. Epidemiol.* **39**, 609–618 (2015).
  - [242] Liu, Q., Chen, L. S., Nicolae, D. L. & Pierce, B. L. A unified set-based test with adaptive filtering for gene-environment interaction analyses. *Biometrics* **72**, 629–38 (2016).
  - [243] Lin, X., Lee, S., Christiani, D. C. & Lin, X. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* **14**, 667–81 (2013).
  - [244] Lin, X. *et al.* Test for rare variants by environment interactions in sequencing association studies. *Biometrics* **72**, 156–64 (2016).
  - [245] Moore, R. *et al.* A linear mixed-model approach to study multivariate geneenvironment interactions. *Nat. Genet.* **1** (2018).
  - [246] Dominicus, A., Skrondal, A., Gjessing, H. K., Pedersen, N. L. & Palmgren, J. Likelihood Ratio Tests in Behavioral Genetics: Problems and Solutions. *Behav. Genet.* **36**, 331–340 (2006).

- [247] Self, S. G. & Liang, K. Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *J. Am. Stat. Assoc.* **82**, 605 (1987).
- [248] Listgarten, J. *et al.* A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics* **29**, 1526–1533 (2013).
- [249] Drikvandi, R., Verbeke, G., Khodadadi, A. & Partovi Nia, V. Testing multiple variance components in linear mixed-effects models. *Biostatistics* **14**, 144–159 (2013).
- [250] Verbeke, G. & Molenberghs, G. The use of score tests for inference on variance components. *Biometrics* **59**, 254–62 (2003).
- [251] Tzeng, J. & Zhang, D. Haplotype-based association analysis via variance-components score test. *Am. J. Hum. Genet.* **81**, 927–38 (2007).
- [252] Greven, S., Crainiceanu, C. M., Küchenhoff, H. & Peters, A. Restricted Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models. *J. Comput. Graph. Stat.* **17**, 870–891 (2008).
- [253] Lippert, C. *et al.* Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. *Bioinformatics* **30**, 3206–14 (2014).
- [254] Kinney, S. K. & Dunson, D. B. Fixed and Random Effects Selection in Linear and Logistic Models. *Biometrics* **63**, 690–698 (2007).
- [255] Chen, Z. & Dunson, D. B. Random effects selection in linear mixed models. *Biometrics* **59**, 762–9 (2003).
- [256] Rasmussen, C. E. Gaussian Processes in Machine Learning. 63–71 (Springer, Berlin, Heidelberg, 2004).
- [257] Searle, S. R. S. R. & Khuri, A. I. Matrix algebra useful for statistics (Wiley, 2017), 2nd edn.
- [258] Duchesne, P. & Lafaye De Micheaux, P. Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Comput. Stat. Data Anal.* **54**, 858–862 (2010).

- [259] Wu, B., Guan, W. & Pankow, J. S. On Efficient and Accurate Calculation of Significance P-Values for Sequence Kernel Association Testing of Variant Set. *Ann. Hum. Genet.* **80**, 123–35 (2016).
- [260] Davies, R. B. Algorithm AS 155: The Distribution of a Linear Combination of  $\chi^2$  Random Variables. *Appl. Stat.* **29**, 323 (1980).
- [261] Liu, H., Tang, Y. & Zhang, H. H. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Comput. Stat. Data Anal.* **53**, 853–856 (2009).
- [262] Pedersen, K. B. P. & S., M. The Matrix Cookbook (Technical University of Denmark, 2012).
- [263] Tipping, M. E. & Bishop, C. M. Probabilistic Principal Component Analysis. *J. R. Stat. Soc.* **61**, 611–622 (1999).
- [264] Mandelbrot, B. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling (Ingram Olkin, Sudhist G. Ghurye, Wassily Hoeffding, William G. Madow, and Henry B. Mann, eds.). *SIAM Rev.* **3**, 80–80 (1961).
- [265] Chen, H., Meigs, J. B. & Dupuis, J. Incorporating Gene-Environment Interaction in Testing for Association with Rare Genetic Variants. *Hum. Hered.* **78**, 81–90 (2014).
- [266] Wang, X., Epstein, M. P. & Tzeng, J. Analysis of Gene-Gene Interactions Using Gene-Trait Similarity Regression. *Hum. Hered.* **78**, 17–26 (2014).
- [267] Zhao, G., Marceau, R., Zhang, D. & Tzeng, J. Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression. *Genetics* **199**, 695–710 (2015).
- [268] Tzeng, J. *et al.* Studying Gene and Gene-Environment Effects of Uncommon and Common Variants on Continuous Traits: A Marker-Set Approach Using Gene-Trait Similarity Regression. *Am. J. Hum. Genet.* **89**, 277–288 (2011).
- [269] Crawford, L., Zeng, P., Mukherjee, S. & Zhou, X. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLOS Genet.* **13**, e1006869 (2017).
- [270] Radhakrishna Rao, C. & Bartlett, M. S. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Math. Proc. Cambridge Philos. Soc.* **44**, 50 (1948).

- [271] Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
- [272] Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **27**, 861–874 (2006).
- [273] Gupta, A. K. A. K. & Nagar, D. K. Matrix variate distributions. 367 (Chapman & Hall, 2000).
- [274] Listgarten, J., Lippert, C. & Heckerman, D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nat. Genet.* **45**, 470–1 (2013).
- [275] Yang, W., Kelly, T. & He, J. Genetic Epidemiology of Obesity. *Epidemiol. Rev.* **29**, 49–61 (2007).
- [276] World Health Organization. Obesity : preventing and managing the global epidemic : report of a WHO consultation. 253 (World Health Organization, 2000).
- [277] Fall, T. & Ingelsson, E. Genome-wide association studies of obesity and metabolic syndrome. *Mol. Cell. Endocrinol.* **382**, 740–757 (2014).
- [278] Adab, P., Pallan, M. & Whincup, P. H. Is BMI the best measure of obesity? *BMJ* **360**, k1274 (2018).
- [279] Sassi, F. Obesity and the Economics of Prevention (Paris: OECD Publishing, 2010).
- [280] Kelly, T., Yang, W., Chen, C. S., Reynolds, K. & He, J. Global burden of obesity in 2005 and projections to 2030. *Int. J. Obes.* (2008).
- [281] Claire Wang, Y., McPherson, K., Marsh, T., Gortmaker, S. L. & Brown, M. Health and economic burden of the projected obesity trends in the USA and the UK. Tech. Rep. (2011).
- [282] Ng, M. *et al.* Global, regional, and national prevalence of overweight and obesity in children and adults during 1980-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* **384**, 766–781 (2014).
- [283] Finucane, M. M. *et al.* National, regional, and global trends in body-mass index since 1980: systematic analysis of health examination surveys and

epidemiological studies with 960 country-years and 9.1 million participants. *Lancet* **377**, 557–567 (2011).

- [284] NCD Risk Factor Collaboration (NCD-RisC). Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19.2 million participants. *Lancet (London, England)* **387**, 1377–1396 (2016).
- [285] Lewis, C. E. *et al.* Mortality, Health Outcomes, and Body Mass Index in the Overweight Range: A Science Advisory From the American Heart Association. *Circulation* **119**, 3263–3271 (2009).
- [286] Haslam, D. W. & James, W. P. T. Obesity. *Lancet* **366**, 1197–1209 (2005).
- [287] Calle, E. E. & Kaaks, R. Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. *Nat. Rev. Cancer* **4**, 579–591 (2004).
- [288] Kurth, T. *et al.* Body mass index and the risk of stroke in men. *Arch. Intern. Med.* **162**, 2557–62 (2002).
- [289] Guh, D. P. *et al.* The incidence of co-morbidities related to obesity and overweight: A systematic review and meta-analysis. *BMC Public Health* **9**, 88 (2009).
- [290] Renehan, A. G., Tyson, M., Egger, M., Heller, R. F. & Zwahlen, M. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *Lancet* **371**, 569–578 (2008).
- [291] Thorpe, K. E., Florence, C. S., Howard, D. H. & Joski, P. Trends: The Impact Of Obesity On Rising Medical Spending. *Health Aff. Suppl Web*, W4–480–6 (2004).
- [292] Stothard, K. J., Tennant, P. W. G., Bell, R. & Rankin, J. Maternal Overweight and Obesity and the Risk of Congenital Anomalies. *JAMA* **301**, 636 (2009).
- [293] Abdelaal, M., le Roux, C. W. & Docherty, N. G. Morbidity and mortality associated with obesity. *Ann. Transl. Med.* **5**, 161 (2017).
- [294] Pi-Sunyer, X. The medical risks of obesity. *Postgrad. Med.* **121**, 21–33 (2009).
- [295] Loos, R. J. & Janssens, A. C. J. Predicting Polygenic Obesity Using Genetic Information. *Cell Metab.* **25**, 535–543 (2017).
- [296] Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41**, 25–34 (2009).



- [297] Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42**, 937–48 (2010).
- [298] Qi, L. & Cho, Y. A. Gene-environment interaction and obesity. *Nutr. Rev.* **66**, 684–94 (2008).
- [299] Swinburn, B. A. *et al.* Series Obesity 1 The global obesity pandemic: shaped by global drivers and local environments. Tech. Rep. (2011).
- [300] Shen, J., Goyal, A. & Sperling, L. The emerging epidemic of obesity, diabetes, and the metabolic syndrome in china. *Cardiol. Res. Pract.* **2012**, 178675 (2012).
- [301] Zaitlen, N. *et al.* Using extended genealogy to estimate components of heritability for 23 quantitative and dichotomous traits. *PLoS Genet.* **9**, e1003520 (2013).
- [302] Maes, H. H., Neale, M. C. & Eaves, L. J. Genetic and environmental factors in relative body weight and human adiposity. *Behav. Genet.* **27**, 325–51 (1997).
- [303] Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat. Genet.* **47**, 1114–1120 (2015).
- [304] Min, J., Chiu, D. T. & Wang, Y. Variation in the heritability of body mass index based on diverse twin studies: a systematic review. *Obes. Rev.* **14**, 871–882 (2013).
- [305] Farooqi, I. S. *et al.* Clinical Spectrum of Obesity and Mutations in the Melanocortin 4 Receptor Gene. *N. Engl. J. Med.* **348**, 1085–1095 (2003).
- [306] Yeo, G. S. *et al.* A frameshift mutation in MC4R associated with dominantly inherited human obesity. *Nat. Genet.* **20**, 111–112 (1998).
- [307] Vaisse, C., Clement, K., Guy-Grand, B. & Froguel, P. A frameshift mutation in human MC4R is associated with a dominant form of obesity. *Nat. Genet.* **20**, 113–114 (1998).
- [308] Krude, H. *et al.* Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by POMC mutations in humans. *Nat. Genet.* **19**, 155–157 (1998).
- [309] Clément, K. *et al.* A mutation in the human leptin receptor gene causes obesity and pituitary dysfunction. *Nature* **392**, 398–401 (1998).

- [310] Montague, C. T. *et al.* Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature* **387**, 903–908 (1997).
- [311] Chung, W. K. An overview of monogenic and syndromic obesities in humans. *Pediatr. Blood Cancer* **58**, 122–8 (2012).
- [312] Goodarzi, M. O. Genetics of obesity: what genetic association studies have taught us about the biology of obesity and its complications. *Lancet Diabetes Endocrinol.* **6**, 223–236 (2018).
- [313] Frayling, T. M. *et al.* A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science* **316**, 889–894 (2007).
- [314] Singh, R. K., Kumar, P. & Mahalingam, K. Molecular genetics of human obesity: A comprehensive review. *C. R. Biol.* **340**, 87–108 (2017).
- [315] Loos, R. J. F. *et al.* Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nat. Genet.* **40**, 768–775 (2008).
- [316] Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
- [317] Yengo, L. *et al.* Meta-analysis of genome-wide association studies for height and body mass index in ~700000 individuals of European ancestry. *Hum. Mol. Genet.* (2018).
- [318] Demerath, E. W. *et al.* The Positive Association of Obesity Variants with Adulthood Adiposity Strengthens over an 80-Year Period: A Gene-by-Birth Year Interaction. *Hum. Hered.* **75**, 175–185 (2013).
- [319] Rosenquist, J. N. *et al.* Cohort of birth modifies the association between FTO genotype and BMI. *Proc. Natl. Acad. Sci.* **112**, 354–359 (2015).
- [320] Rask-Andersen, M., Karlsson, T., Ek, W. E. & Johansson, Å. Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. *PLOS Genet.* **13**, e1006977 (2017).
- [321] Johnson, W. *et al.* Modification of genetic influences on adiposity between 36 and 63 years of age by physical activity and smoking in the 1946 British Birth Cohort Study. *Nutr. Diabetes* **4**, e136–e136 (2014).

- [322] Zhu, J. *et al.* Associations of Genetic Risk Score with Obesity and Related Traits and the Modifying Effect of Physical Activity in a Chinese Han Population. *PLoS One* **9**, e91442 (2014).
- [323] Qi, Q. *et al.* Television Watching, Leisure Time Physical Activity, and the Genetic Predisposition in Relation to Body Mass Index in Women and Men. *Circulation* **126**, 1821–1827 (2012).
- [324] Richardson, A. S. *et al.* Moderate to vigorous physical activity interactions with genetic variants and body mass index in a large US ethnically diverse cohort. *Pediatr. Obes.* **9**, e35–46 (2014).
- [325] Qi, Q. *et al.* Sugar-sweetened beverages and genetic risk of obesity. *N. Engl. J. Med.* **367**, 1387–96 (2012).
- [326] Liu, H. & Guo, G. Lifetime Socioeconomic Status, Historical Context, and Genetic Inheritance in Shaping Body Mass in Middle and Late Adulthood. *Am. Sociol. Rev.* **80**, 705–737 (2015).
- [327] Ng, M. C. Y. *et al.* Discovery and fine-mapping of adiposity loci using high density imputation of genome-wide association studies in individuals of African ancestry: African Ancestry Anthropometry Genetics Consortium. *PLOS Genet.* **13**, e1006719 (2017).
- [328] Wen, W. *et al.* Meta-analysis of genome-wide association studies in East Asian-ancestry populations identifies four new loci for body mass index. *Hum. Mol. Genet.* **23**, 5492–5504 (2014).
- [329] Wang, T. *et al.* Effects of Obesity Related Genetic Variations on Visceral and Subcutaneous Fat Distribution in a Chinese Population. *Sci. Rep.* **6**, 20691 (2016).
- [330] Elks, C. E. *et al.* Adult obesity susceptibility variants are associated with greater childhood weight gain and a faster tempo of growth: the 1946 British Birth Cohort Study. *Am. J. Clin. Nutr.* **95**, 1150–1156 (2012).
- [331] Ahmad, S. *et al.* Established BMI-associated genetic variants and their prospective associations with BMI and other cardiometabolic traits: the GLACIER Study. *Int. J. Obes.* **40**, 1346–1352 (2016).
- [332] Kostem, E. & Eskin, E. Improving the Accuracy and Efficiency of Partitioning Heritability into the Contributions of Genomic Regions. *Am. J. Hum. Genet.* **92**, 558–564 (2013).

- [333] Kass, R. E. & Raftery, A. E. Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
- [334] Clark, S. A. & van der Werf, J. Genomic Best Linear Unbiased Prediction (gBLUP) for the Estimation of Genomic Breeding Values. In *Methods Mol. Biol.*, vol. 1019, 321–330 (2013).
- [335] Lee, S. H., van der Werf, J. H. J., Hayes, B. J., Goddard, M. E. & Visscher, P. M. Predicting Unobserved Phenotypes for Complex Traits from Whole-Genome SNP Data. *PLoS Genet.* **4**, e1000231 (2008).
- [336] Henderson, C. R. Applications of linear models in animal breeding. 462 (University of Guelph, 1984).
- [337] Gumedze, F. N. & Dunne, T. T. Parameter estimation and inference in the linear mixed model. *Linear Algebra Appl.* **435**, 1920–1944 (2011).
- [338] Fesinmeyer, M. D. *et al.* Genetic Risk Factors for BMI and Obesity in an Ethnically Diverse Population: Results from the Population Architecture Using Genomics and Epidemiology (PAGE) Study. *Obesity* **21**, 835–846 (2013).
- [339] Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
- [340] Almli, L. M. *et al.* Correcting Systematic Inflation in Genetic Association Tests That Consider Interaction Effects. *JAMA Psychiatry* **71**, 1392 (2014).
- [341] Voorman, A., Lumley, T., McKnight, B. & Rice, K. Behavior of QQ-Plots and Genomic Control in Studies of Gene-Environment Interaction. *PLoS One* **6**, e19416 (2011).
- [342] Tchetgen Tchetgen, E. J. & Kraft, P. On the Robustness of Tests of Genetic Associations Incorporating Gene-environment Interaction When the Environmental Exposure is Misspecified. *Epidemiology* **22**, 257–261 (2011).
- [343] Cornelis, M. C. *et al.* Gene-Environment Interactions in Genome-Wide Association Studies: A Comparative Study of Tests Applied to Empirical Studies of Type 2 Diabetes. *Am. J. Epidemiol.* **175**, 191–202 (2012).
- [344] Rao, T. J. & Province, M. A. A Framework for Interpreting Type I Error Rates from a Product-Term Model of Interaction Applied to Quantitative Traits. *Genet. Epidemiol.* **40**, 144–53 (2016).

- [345] Lindgren, C. M. *et al.* Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. *PLoS Genet.* **5**, e1000508 (2009).
- [346] Graff, M. *et al.* Generalization of adiposity genetic loci to US Hispanic women. *Nutr. Diabetes* **3**, e85 (2013).
- [347] Scherag, A. *et al.* Two new Loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and german study groups. *PLoS Genet.* **6**, e1000916 (2010).
- [348] Warrington, N. M. *et al.* Modelling BMI Trajectories in Children for Genetic Association Studies. *PLoS One* **8**, e53897 (2013).
- [349] Hall, N. G., Klenotic, P., Anand-Apte, B. & Apte, S. S. ADAMTSL-3/punctin-2, a novel glycoprotein in extracellular matrix related to the ADAMTS family of metalloproteases. *Matrix Biol.* **22**, 501–10 (2003).
- [350] Zillikens, M. C. *et al.* Large meta-analysis of genome-wide association studies identifies five loci for lean body mass. *Nat. Commun.* **8**, 80 (2017).
- [351] Wen, W. *et al.* Genome-wide association studies in East Asians identify new loci for waist-hip ratio and waist circumference. *Sci. Rep.* **6**, 17958 (2016).
- [352] Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
- [353] Dastani, Z. *et al.* Novel Loci for Adiponectin Levels and Their Influence on Type 2 Diabetes and Metabolic Traits: A Multi-Ethnic Meta-Analysis of 45,891 Individuals. *PLoS Genet.* **8**, e1002607 (2012).
- [354] Wu, Y. *et al.* A meta-analysis of genome-wide association studies for adiponectin levels in East Asians identifies a novel locus near WDR11-FGFR2. *Hum. Mol. Genet.* **23**, 1108–1119 (2014).
- [355] Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44**, 659–669 (2012).
- [356] Willer, C. J. *et al.* Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).

- [357] Kathiresan, S. *et al.* A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med. Genet.* **8**, S17 (2007).
- [358] Cho, Y. S. *et al.* Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat. Genet.* **44**, 67–72 (2012).
- [359] Fox, C. S. *et al.* Genome-wide association to body mass index and waist circumference: the Framingham Heart Study 100K project. *BMC Med. Genet.* **8**, S18 (2007).
- [360] Xiao, R. & Boehnke, M. Quantifying and correcting for the winner’s curse in genetic association studies. *Genet. Epidemiol.* **33**, 453–62 (2009).
- [361] Dahl, A. *et al.* A multiple-phenotype imputation method for genetic studies. *Nat. Genet.* **48**, 466–472 (2016).
- [362] Little, R. J. A. & Rubin, D. B. Statistical analysis with missing data. 278 (Wiley, 1987).
- [363] Freimer, N. & Sabatti, C. The Human Phenome Project. *Nat. Genet.* **34**, 15–21 (2003).
- [364] Jones, R., Pembrey, M., Golding, J. & Herrick, D. The search for genotype/phenotype associations and the phenome scan. *Paediatr. Perinat. Epidemiol.* **19**, 264–275 (2005).
- [365] Bilder, R. *et al.* Phenomics: the systematic study of phenotypes on a genome-wide scale. *Neuroscience* **164**, 30–42 (2009).
- [366] Ghebranious, N., McCarty, C. A. & Wilke, R. A. Clinical phenome scanning. *Per. Med.* **4**, 175–182 (2007).
- [367] Robinson, J. R., Denny, J. C., Roden, D. M. & Van Driest, S. L. Genome-wide and Phenome-wide Approaches to Understand Variable Drug Actions in Electronic Health Records. *Clin. Transl. Sci.* **11**, 112–122 (2018).
- [368] McCarty, C. A. *et al.* The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med. Genomics* **4**, 13 (2011).
- [369] Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31**, 1102–1111 (2013).

- [370] Panagiotou, O. A. *et al.* A genome-wide pleiotropy scan for prostate cancer risk. *Eur. Urol.* **67**, 649–57 (2015).
- [371] Campa, D. *et al.* A Genome-Wide Pleiotropy Scan Does Not Identify New Susceptibility Loci for Estrogen Receptor Negative Breast Cancer. *PLoS One* **9**, e85955 (2014).
- [372] Pierce, B. L. & Ahsan, H. Genome-Wide "Pleiotropy Scan" Identifies HNF1A Region as a Novel Pancreatic Cancer Susceptibility Locus. *Cancer Res.* **71**, 4352–4358 (2011).
- [373] Kocarnik, J. M. *et al.* Pleiotropic and Sex-Specific Effects of Cancer GWAS SNPs on Melanoma Risk in the Population Architecture Using Genomics and Epidemiology (PAGE) Study. *PLoS One* **10**, e0120491 (2015).
- [374] Park, S. L. *et al.* Association of Cancer Susceptibility Variants with Risk of Multiple Primary Cancers: The Population Architecture using Genomics and Epidemiology Study. *Cancer Epidemiol. Biomarkers Prev.* **23**, 2568–2578 (2014).
- [375] Setiawan, V. W. *et al.* Cross-cancer pleiotropic analysis of endometrial cancer: PAGE and E2C2 consortia. *Carcinogenesis* **35**, 2068–73 (2014).
- [376] Park, S. L. *et al.* Pleiotropic Associations of Risk Variants Identified for Other Cancers With Lung Cancer Risk: The PAGE and TRICL Consortia. *JNCI J. Natl. Cancer Inst.* **106**, dju061 (2014).
- [377] Cheng, I. *et al.* Pleiotropic effects of genetic risk variants for other cancers on colorectal cancer risk: PAGE, GECCO and CCFR consortia. *Gut* **63**, 800–807 (2014).
- [378] Cotsapas, C. *et al.* Pervasive Sharing of Genetic Effects in Autoimmune Disease. *PLoS Genet.* **7**, e1002254 (2011).
- [379] Sanseau, P. *et al.* Use of genome-wide association studies for drug repositioning. *Nat. Biotechnol.* **30**, 317–320 (2012).
- [380] Rastegar-Mojarad, M., Ye, Z., Kolesar, J. M., Hebbring, S. J. & Lin, S. M. Opportunities for drug repositioning from phenome-wide association studies. *Nat. Biotechnol.* **33**, 342–345 (2015).
- [381] Diogo, D. *et al.* Phenome-wide association studies across large population cohorts support drug target validation. *Nat. Commun.* **9**, 4285 (2018).

- [382] Millard, L. A., Davies, N. M., Gaunt, T. R., Davey Smith, G. & Tilling, K. Software Application Profile: PHESANT: a tool for performing automated phenome scans in UK Biobank. *Int. J. Epidemiol.* **47**, 29–35 (2018).
- [383] Staley, J. R. *et al.* PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).
- [384] Pendergrass, S. A. *et al.* The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet. Epidemiol.* **35**, 410–22 (2011).
- [385] Liao, K. P. *et al.* Phenome-Wide Association Study of Autoantibodies to Citrullinated and Noncitrullinated Epitopes in Rheumatoid Arthritis. *Arthritis Rheumatol. (Hoboken, N.J.)* **69**, 742–749 (2017).
- [386] Shameer, K. *et al.* A genome- and phenome-wide association study to identify genetic variants influencing platelet count and volume and their pleiotropic effects. *Hum. Genet.* **133**, 95–109 (2014).
- [387] Verma, S. S. *et al.* Phenome-Wide Interaction Study (PheWIS) In AIDS Clinical Trials Group Data (ACTG). *Pac. Symp. Biocomput.* **21**, 57–68 (2016).
- [388] Robinson, M. R. *et al.* Genotype-covariate interaction effects and the heritability of adult body mass index. *Nat. Genet.* **49**, 1174–1181 (2017).
- [389] Mayhew, A. J. & Meyre, D. Assessing the Heritability of Complex Traits in Humans: Methodological Challenges and Opportunities. *Curr. Genomics* **18**, 332–340 (2017).
- [390] Sandoval-Motta, S., Aldana, M., Martínez-Romero, E. & Frank, A. The Human Microbiome and the Missing Heritability Problem. *Front. Genet.* **8**, 80 (2017).
- [391] Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci.* **109**, 1193–1198 (2012).
- [392] Zheng, J. *et al.* Genome-Wide Contribution of Genotype by Environment Interaction to Variation of Diabetes-Related Traits. *PLoS One* **8**, e77442 (2013).
- [393] Liley, J. & Wallace, C. A Pleiotropy-Informed Bayesian False Discovery Rate Adapted to a Shared Control Design Finds New Disease Associations From GWAS Summary Statistics. *PLOS Genet.* **11**, e1004926 (2015).



- [394] Andreassen, O. A. *et al.* Improved Detection of Common Variants Associated with Schizophrenia and Bipolar Disorder Using Pleiotropy-Informed Conditional False Discovery Rate. *PLoS Genet.* **9**, e1003455 (2013).
- [395] Ignatiadis, N., Klaus, B., Zaugg, J. B. & Huber, W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods* **13**, 577–580 (2016).
- [396] Sun, L., Craiu, R. V., Paterson, A. D. & Bull, S. B. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genet. Epidemiol.* **30**, 519–530 (2006).
- [397] Schork, A. J. *et al.* All SNPs Are Not Created Equal: Genome-Wide Association Studies Reveal a Consistent Pattern of Enrichment among Functionally Annotated SNPs. *PLoS Genet.* **9**, e1003449 (2013).
- [398] Pain, O., Dudbridge, F. & Ronald, A. Are your covariates under control? How normalization can re-introduce covariate effects. *Eur. J. Hum. Genet.* **26**, 1194–1201 (2018).
- [399] McInnes, G. *et al.* Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics. *bioRxiv* 304188 (2018).
- [400] Golan, D., Lander, E. S. & Rosset, S. Measuring missing heritability: inferring the contribution of common variants. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E5272–81 (2014).
- [401] Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–8 (2010).
- [402] He, M. *et al.* Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. *Hum. Mol. Genet.* **24**, 1791–800 (2015).
- [403] Hwang, J. *et al.* Genome-wide association meta-analysis identifies novel variants associated with fasting plasma glucose in East Asians. *Diabetes* **64**, 291–8 (2015).
- [404] Bhardwaj, G., Penniman, C. M., Suarez Beltran, P. A., Foster, C. M. & O'Neill, B. T. Loss of Insulin/IGF-1 Receptors in Muscle Coordinately Downregulates Mitochondrial Metabolism and Alters Mitophagy via Foxo Transcription Factors. *Diabetes* **67**, 187–OR (2018).
- [405] Boucher, J. *et al.* Differential Roles of Insulin and IGF-1 Receptors in Adipose Tissue Development and Function. *Diabetes* **65**, 2201–13 (2016).

- [406] Kineman, R. D., del Rio-Moreno, M. & Sarmiento-Cabral, A. 40 YEARS of IGF1: Understanding the tissue-specific roles of IGF1/IGF1R in regulating metabolism using the Cre/loxP system. *J. Mol. Endocrinol.* **61**, T187–T198 (2018).
- [407] Klöting, N. *et al.* Autocrine IGF-1 action in adipocytes controls systemic IGF-1 concentrations and growth. *Diabetes* **57**, 2074–82 (2008).
- [408] Spadaro, O. *et al.* IGF1 Shapes Macrophage Activation in Response to Immunometabolic Challenge. *Cell Rep.* **19**, 225–234 (2017).
- [409] Müller, K. *et al.* TSH Compensates Thyroid-Specific IGF-I Receptor Knockout and Causes Papillary Thyroid Hyperplasia. *Mol. Endocrinol.* **25**, 1867–1879 (2011).
- [410] Simino, J. *et al.* Gene-age interactions in blood pressure regulation: a large-scale investigation with the CHARGE, Global BPgen, and ICBP Consortia. *Am. J. Hum. Genet.* **95**, 24–38 (2014).
- [411] Newton-Cheh, C. *et al.* Genome-wide association study identifies eight loci associated with blood pressure. *Nat. Genet.* **41**, 666–76 (2009).
- [412] Hirokawa, M. *et al.* A genome-wide association study identifies PLCL2 and AP3D1-DOT1L-SF3A2 as new susceptibility loci for myocardial infarction in Japanese. *Eur. J. Hum. Genet.* **23**, 374–80 (2015).
- [413] Kato, N. *et al.* Trans-ancestry genome-wide association study identifies 12 genetic loci influencing blood pressure and implicates a role for DNA methylation. *Nat. Genet.* **47**, 1282–1293 (2015).
- [414] Liu, C. *et al.* Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. *Nat. Genet.* **48**, 1162–1170 (2016).
- [415] Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41**, 35–46 (2009).
- [416] Kato, N. *et al.* High-density association study and nomination of susceptibility genes for hypertension in the Japanese National Project. *Hum. Mol. Genet.* **17**, 617–627 (2007).
- [417] Wang, Y. *et al.* Whole-genome association study identifies STK39 as a hypertension susceptibility gene. *Proc. Natl. Acad. Sci.* **106**, 226–231 (2009).

- [418] Adeyemo, A. *et al.* A genome-wide association study of hypertension and blood pressure in African Americans. *PLoS Genet.* **5**, e1000564 (2009).
- [419] Chen, T. *et al.* Basal metabolic rate is correlated with blood pressure among young population. *Int. J. Clin. Exp. Med.* (2016).
- [420] Snodgrass, J. J., Leonard, W. R., Sorensen, M. V., Tarskaia, L. A. & Mosher, M. The influence of basal metabolic rate on blood pressure among indigenous Siberians. *Am. J. Phys. Anthropol.* **137**, 145–155 (2008).
- [421] Ali, N. *et al.* Hypertension prevalence and influence of basal metabolic rate on blood pressure among adult students in Bangladesh. *BMC Public Health* **18**, 58 (2017).
- [422] Speed, D., Hemani, G., Johnson, M. R. & Balding, D. J. Improved heritability estimation from genome-wide SNPs. *Am. J. Hum. Genet.* **91**, 1011–21 (2012).
- [423] Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
- [424] Speed, D. *et al.* Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).
- [425] Speed, D. & Balding, D. J. MultiBLUP: improved SNP-based prediction for complex traits. *Genome Res.* **24**, 1550–7 (2014).
- [426] Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
- [427] Garrod, A. E. The incidence of alkaptonuria: a study in chemical individuality. 1902. *Mol. Med.* **2**, 274–82 (1996).
- [428] Noyce, A. J. *et al.* Estimating the causal influence of body mass index on risk of Parkinson disease: A Mendelian randomisation study. *PLOS Med.* **14**, e1002314 (2017).
- [429] Wauer, T., Simicek, M., Schubert, A. & Komander, D. Mechanism of phospho-ubiquitin-induced PARKIN activation. *Nature* **524**, 370–374 (2015).
- [430] Narendra, D. P. & Youle, R. J. Targeting Mitochondrial Dysfunction: Role for PINK1 and Parkin in Mitochondrial Quality Control. *Antioxid. Redox Signal.* **14**, 1929–1938 (2011).

- [431] Lazarou, M. *et al.* The ubiquitin kinase PINK1 recruits autophagy receptors to induce mitophagy. *Nature* **524**, 309–314 (2015).
- [432] Koyano, F. *et al.* Ubiquitin is phosphorylated by PINK1 to activate parkin. *Nature* **510**, 162–166 (2014).
- [433] Kane, L. A. *et al.* PINK1 phosphorylates ubiquitin to activate Parkin E3 ubiquitin ligase activity. *J. Cell Biol.* **205**, 143–153 (2014).
- [434] Costa, D. K. *et al.* Reduced intestinal lipid absorption and body weight-independent improvements in insulin sensitivity in high-fat diet-fed Park2 knockout mice. *Am. J. Physiol. Endocrinol. Metab.* **311**, E105–16 (2016).
- [435] Kim, K. *et al.* Parkin is a lipid-responsive regulator of fat uptake in mice and mutant human cells. *J. Clin. Invest.* **121**, 3701–12 (2011).
- [436] Young, A. I., Wauthier, F. L. & Donnelly, P. Identifying loci affecting trait variability and detecting interactions in genome-wide association studies. *Nat. Genet.* **50**, 1608–1614 (2018).
- [437] Gao, C. *et al.* Model-based and Model-free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson’s Disease. *Sci. Rep.* **8**, 7129 (2018).

Appendix A: Copy of manuscript  
entitled, ‘A linear mixed-model  
approach to study multivariate  
gene-environment interactions’,  
published by Nature Genetics

# A linear mixed-model approach to study multivariate gene–environment interactions

Rachel Moore<sup>1,2,3,9</sup>, Francesco Paolo Casale<sup>4,9</sup>, Marc Jan Bonder<sup>2</sup>, Danilo Horta<sup>2</sup>, BIOS Consortium<sup>5</sup>, Lude Franke<sup>6</sup>, Inês Barroso<sup>1\*</sup> and Oliver Stegle<sup>1,2,7,8\*</sup>

**Different exposures, including diet, physical activity, or external conditions can contribute to genotype–environment interactions (G×E). Although high-dimensional environmental data are increasingly available and multiple exposures have been implicated with G×E at the same loci, multi-environment tests for G×E are not established. Here, we propose the structured linear mixed model (StructLMM), a computationally efficient method to identify and characterize loci that interact with one or more environments. After validating our model using simulations, we applied StructLMM to body mass index in the UK Biobank, where our model yields previously known and novel G×E signals. Finally, in an application to a large blood eQTL dataset, we demonstrate that StructLMM can be used to study interactions with hundreds of environmental variables.**

Large population cohorts that combine genetic profiling with deep phenotype and environmental data, including diet, physical activity and other lifestyle covariates, have fostered interest to study G×E. Already, such analyses have identified G×E for different traits in humans, including disease risk<sup>1,2</sup> and molecular traits<sup>3,4</sup>.

Established G×E methods test for interactions between a single environmental variable and individual genetic variants<sup>5</sup>. Recent extensions enable assessment of G×E across sets of genetic variants, either using genetic risk scores<sup>6</sup> or variance component tests<sup>7–9</sup>. Although there is evidence that multiple environments can interact with a single genetic locus to influence phenotypes (for example, a number of environments have been shown to alter the effect of *FTO* on body mass index (BMI), including physical activity<sup>10–13</sup>, diet<sup>12–15</sup> and smoking<sup>12</sup>), there are no robust methods for the joint G×E analysis of multiple environmental variables. Multivariate G×E tests can have power advantages, in particular to identify interactions that are simultaneously driven by multiple environments or because combinations of multiple environmental variables act as proxy for unobserved drivers of G×E. Additionally, joint tests reduce the multiple testing burden. Thus, as increasingly high-dimensional environmental data are available in population cohorts and given the desire to fully understand the impact of multiple environments in complex traits and diseases, there is a growing need for multi-environment G×E tests. Here, we present StructLMM, a variance component test to identify and characterize G×E interactions with multiple environments. Our model can handle hundreds of environmental variables, and it can be applied to large cohorts of hundreds of thousands of individuals.

## Results

Conventional linear mixed models (LMMs) are used to test for associations with constant genetic effect sizes across individuals in the population, also called persistent genetic effects. Covariates and additional random effect components are included to account for

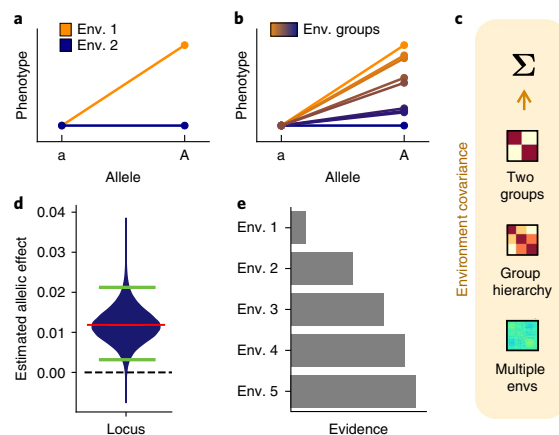
population structure, environment, or other additive (confounding) effects. StructLMM extends the LMM framework by modeling heterogeneity in effect sizes due to G×E

$$y = \underbrace{Xb}_{\text{covariates}} + \underbrace{x\beta_G}_{\text{persistent G}} + \underbrace{x\odot\beta_{G\times E}}_{\text{G}\times\text{E}} + \underbrace{e}_{\text{environment}} + \underbrace{\psi}_{\text{noise}} \quad (1)$$

Here,  $\beta_G$  denotes the effect size of a conventional persistent genetic effect component, and  $\beta_{G\times E} = [\beta_{G\times E}^1, \dots, \beta_{G\times E}^N]^T$  is a vector of per-individual effect sizes to account for heterogeneous genetic effects, which follows a multivariate normal distribution,  $\beta_{G\times E} \sim N(0, \sigma_{G\times E}^2 \Sigma)$ . Depending on the functional form of the environmental covariance  $\Sigma$ , this model can account for different types of G×E, for example, hierarchies of discrete environmental groups, or as considered here, G×E effects based on a set of continuous and discrete environmental covariates (Fig. 1b,c). The same environmental covariance is also used to account for additive environmental effects,  $e \sim N(0, \Sigma)$ . StructLMM is technically related to existing variance component tests for rare variants<sup>16</sup> and epistasis<sup>17</sup> (comparison to alternative methods in Supplementary Note).

Using the multi-environment model defined above (equation (1)), we propose a score test to identify loci with significant G×E interaction effects. Additionally, the same framework can be used to define a joint association test that accounts for the possibility of heterogeneous effect sizes due to G×E, which generalizes previous two degrees of freedom single-environment association tests<sup>5,18</sup>. Both tests are computationally efficient, enabling genome-wide analyses using hundreds of environmental variables on cohorts of hundreds of thousands of individuals. The model facilitates different analyses to characterize G×E effects at individual loci, including estimation of the fraction of genetic variance explained by G×E ( $\rho$ , Methods) and estimating per-individual allelic effects based on environmental profiles in the population (Fig. 1d), thus identifying individuals

<sup>1</sup>Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. <sup>2</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. <sup>3</sup>University of Cambridge, Cambridge, UK. <sup>4</sup>Microsoft Research New England, Cambridge, Massachusetts, USA. <sup>5</sup>A full list of members and affiliations appears at the end of the paper. <sup>6</sup>University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands. <sup>7</sup>European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. <sup>8</sup>Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. <sup>9</sup>These authors contributed equally: Rachel Moore, Francesco Paolo Casale. \*e-mail: [ib1@sanger.ac.uk](mailto:ib1@sanger.ac.uk); [oliver.stegle@embl.de](mailto:oliver.stegle@embl.de)



**Fig. 1 | Overview of the StructLMM model.** **a**, Basic genotype–environment interaction, with a genetic effect that is specific to one of two groups (blue and orange lines correspond to the average phenotypes observed within two environmental groups for two alleles). Env, environment. **b**, Interaction with multiple environmental groups or bins of continuous environmental states (average phenotypes for groups exerting increasing G×E effects, from blue to orange, for two alleles). **c**, StructLMM accounts for possible heterogeneity in effect sizes due to G×E using a multivariate normal prior, where alternative choices of the environmental covariance  $\Sigma$  can capture discrete (two groups, group hierarchy (**a,b**)) or continuous substructure of environmental exposures in the population (multiple environments). **d,e**, Different illustrative example analyses using StructLMM. **d**, Estimation of per-individual allelic effects in the population at individual loci. The violin plot displays the density of estimated allelic effect sizes for individuals in the population. Median and the top and bottom 5% quantiles of the effect size distribution are indicated by the red and green bars, respectively. **e**, Bayes factors between the full model and models with environmental variables removed, thereby identifying environments that are most relevant for G×E.

at increased or decreased trait risk. Finally, StructLMM can be used to explore which environments are most relevant for G×E by comparing models that contain all environmental factors and models with environmental variables removed (Fig. 1e; full derivation in Methods).

**Model validation using simulated data.** Initially, we considered simulated data using genotypes from the 1000 Genomes project<sup>19</sup> to assess the statistical calibration and power of StructLMM. To mimic environmental distributions as observed in real settings, we simulated G×E based on 60 environmental covariates from UK Biobank, including physical activity, diet, and other lifestyle factors (Methods). We varied the sample size of the simulated population, the magnitude of G×E effects, the number of driving environments for G×E, and other parameters (Supplementary Table 1).

First, we confirmed the statistical calibration of the StructLMM interaction test (referred to as StructLMM-int), either considering phenotypes simulated without any genetic effects (Fig. 2a and Supplementary Figure 1a,b) or simulating from a persistent effect model without interactions (the null model of StructLMM-int; Supplementary Figure 1a,b).

Next, we simulated phenotypes with variable fractions of the genetic variance explained by G×E ( $\rho$ , Methods) and assessed the power of StructLMM-int. For comparison, we also considered a single-environment one degree of freedom fixed effect test (SingleEnv-Renv-int, Supplementary Table 2; as described in Gauderman et al.<sup>18</sup>,

Bonferroni adjusted for the number of environments, Methods), using the same random effect component (as for StructLMM) to account for additive environmental effects under the null.

The power of both tests increases as the fraction of the genetic effect explained by G×E ( $\rho$ ) increases, noting that StructLMM-int is substantially better powered than the SingleEnv-Renv-int test (Fig. 2b and Supplementary Fig. 2a). As a second parameter, we varied the number of active environments that contribute to G×E but used all 60 environmental variables during testing. The results of this analysis show that StructLMM-int increasingly outperforms the corresponding SingleEnv-Renv-int G×E test as the number of active environments increases (Fig. 2c and Supplementary Fig. 2b).

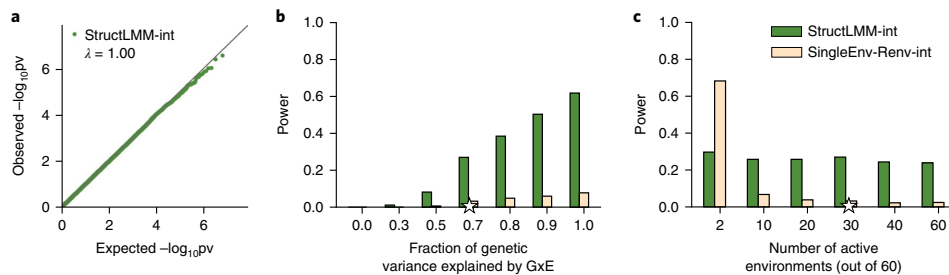
We considered a number of additional settings, including varying the total number of observed environments, the fraction of phenotypic variance explained by additive environmental effects and simulating interaction effects using environments that are not included at the testing stage, with the latter corresponding to G×E effects driven by environments for which there are no measurements available, a scenario that is likely to occur in practice. We also considered settings where the environments were themselves heritable, varied the extent of distributional skew and considered binary environments with different frequencies. Across all settings, StructLMM-int had consistent power advantages over alternative methods and remained calibrated (Supplementary Figs. 2–4).

For the same settings, we also considered the StructLMM joint association test, which accounts for the possibility of heterogeneous effect sizes due to G×E, and compared it to a two degrees of freedom single-environment test using fixed effects (as described in Kraft et al.<sup>5</sup>; Bonferroni adjusted for the number of environments, Methods; Supplementary Table 2), as well as to conventional association tests that only model persistent effects (Supplementary Table 2). In these experiments (Supplementary Figs. 2–4), the StructLMM joint association test yielded similar power advantages as StructLMM-int when testing for interactions, indicating that StructLMM can be useful to discover additional associations, in particular for variants with strong G×E (Supplementary Fig. 2a).

Finally, we considered alternative implementations of interaction and association tests (Supplementary Table 2), using fixed effects to account for additive environment instead of a random effect component, which yielded near-identical results (Supplementary Figure 2). We also note that multi-environment G×E tests can in principle be implemented based on fixed effect tests with as many degrees of freedoms as environments (Supplementary Table 2). However, we observed that such tests were not always calibrated (Supplementary Figure 1b), in particular for large numbers of environments, and in addition had lower performance (Supplementary Fig. 1c,d).

Taken together, these results show increased power and robustness of StructLMM compared with existing methods, in particular when large numbers of environments drive G×E interaction effects, as might be expected to occur for the majority of complex traits and diseases.

**Application to data from UK Biobank.** Initially, we applied StructLMM-int to test for G×E interactions at 97 variants (corresponding genes as annotated by GIANT<sup>20</sup>) that have previously been linked to BMI using independent data<sup>20</sup>. We considered 252,188 unrelated individuals of European ancestry, for which BMI and 64 lifestyle covariates, similar to those used in Young et al.<sup>13</sup> (12 diet-related factors, three factors linked to physical activity and six lifestyle factors, modeled as gender adjusted and age adjusted; Methods and Supplementary Figures 5 and 6), were available in the full release of UK Biobank<sup>21</sup>. StructLMM-int identified four significant G×E effects ( $\alpha < 0.05$ , Bonferroni adjusted), whereas a single-environment one degree of freedom fixed-effect test (SingleEnv-Renv-int), identified only two of these interactions (Fig. 3a, Supplementary Fig. 7 and



**Fig. 2 | Assessment of statistical calibration and power using simulated data.** **a**, QQ plots of negative log  $P$  values from the StructLMM interaction test (green, StructLMM-int) using phenotypes simulated from the null (no genetic effect) for 103,527 variants on chromosome 21. **b**, Comparison of power for detecting G×E interactions for increasing fractions of the genetic variance explained by G×E ( $\rho$ ), comparing the StructLMM interaction test (StructLMM-int) and a single-environment interaction test (SingleEnv-Renv-int). **c**, Analogous power analysis, when simulating G×E using increasing numbers of active environments with non-zero G×E effects (out of 60 environments total, considered in all tests;  $\rho=0.7$ ). All 60 environments contribute to the simulated additive environment effect. Models were assessed in terms of power (at Family Wise Error Rate, FWER < 1%) for detecting variants with true G×E effects (Methods). Stars denote default values of genetic parameters, which were retained when varying other parameters (Supplementary Table 1). A synthetic European population of 5,000 individuals based on the 1000 Genomes Project was used for all experiments.

Supplementary Table 3). Among the loci identified by StructLMM-int was the *FTO* locus (rs1558902,  $\rho=0.14$ ; Supplementary Fig. 8a), which has previously been implicated in G×E for multiple environments<sup>10,12–14</sup>; *MC4R* (Fig. 3b), for which an interaction with physical activity in females aged 20–40 years has been suggested previously ( $P_{\text{adj}}=0.025$ , reported in ref. <sup>12</sup>); *SEC16B* (Supplementary Fig. 8b), for which secondary analyses provided some evidence for an interaction ( $P=0.025$ ) with physical activity in Europeans<sup>11</sup> and in a separate study in Hispanics<sup>22</sup>; and *PARK2* (Supplementary Fig. 8c), a gene that has been linked to time-dependent variation in BMI<sup>23</sup>. StructLMM also enhanced the significance of interactions identified by both tests ( $P_{\text{StructLMM-int}}=4.23 \times 10^{-16}$  versus  $P_{\text{adj SingleEnv-Renv-int}}=6.76 \times 10^{-6}$  and  $P_{\text{StructLMM-int}}=1.15 \times 10^{-4}$  versus  $P_{\text{adj SingleEnv-Renv-int}}=4.48 \times 10^{-4}$  for *FTO* and *SEC16B*, respectively). Larger differences in the number of discoveries were observed at more lenient thresholds, for example, 11 versus six loci with G×E at false discovery rate (FDR) < 5% (Benjamini–Hochberg adjustment; Supplementary Table 3).

We also considered additional fixed effect interaction tests, including multi-environment G×E tests based on fixed effects, which identified fewer interactions than StructLMM-int ( $N=2$  versus  $N=4$ ;  $\alpha < 0.05$ ), as well as alternative implementations of the single-environment interaction test, which consistent with the results on simulations, yielded near-identical results to SingleEnv-Renv-int (Supplementary Figs. 9 and 10 and Supplementary Table 3).

Finally, as an alternative filtering strategy, we applied the same interaction tests to 17,606 variants with significant persistent associations with BMI in UK Biobank ( $P < 5 \times 10^{-8}$ ; LMM-Renv). StructLMM-int identified 23 loci with G×E interactions (FDR < 5%; Benjamini–Hochberg adjusted,  $\pm 500$  kb,  $r^2 > 0.1$ ), including *SEC16B*, *MC4R* and *FTO*, compared with, at most, 11 loci identified by alternative methods (Supplementary Fig. 11 and Supplementary Table 3).

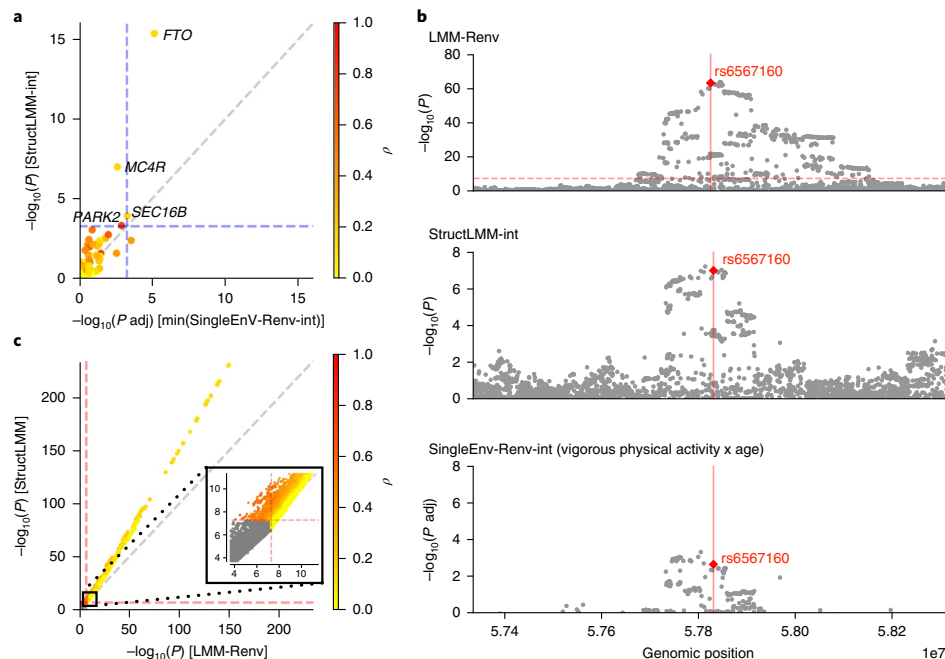
The StructLMM framework can also be used to test for associations while accounting for the possibility of effect size heterogeneity due to G×E. To explore this option, we applied the StructLMM joint association test to BMI, using low-frequency and common variants (imputed variants, minor allele frequency (MAF) > 1%, 7,515,856 variants in total) and the same set of 64 lifestyle covariates as considered in the interaction analysis. For comparison, we also considered an LMM using the same random effect component to account for additive environmental effects, as in StructLMM (LMM-Renv), and a linear model without accounting for additive environment (LM).

Although the choice of null model can have a large impact on loci discovery ( $P < 5 \times 10^{-8}$ ,  $\pm 500$  kb,  $r^2 > 0.1$ ; LMM-Renv: 327 loci, of which 14.37% were not detected by the LM; LM, 379 loci, of which 25.59% were not detected by LMM-Renv; Supplementary Table 4 and Supplementary Fig. 12), StructLMM identified 23 loci that were not detected by other methods (351 unique loci in total; Fig. 3c, Supplementary Table 4 and Supplementary Figs. 13–16), indicating that the StructLMM joint association test can be used to identify additional loci with a strong G×E component. One such locus lies in the *ADAMTSL3* gene (rs4842838,  $P_{\text{StructLMM}}=9.35 \times 10^{-10}$ ,  $P_{\text{LMM}}=3.83 \times 10^{-7}$ ,  $P_{\text{LM}}=2.37 \times 10^{-5}$ ), which codes for a glycoprotein<sup>24</sup>. Other variants within this gene have been linked to BMI-related traits, including lean body mass<sup>25</sup>, waist circumference<sup>26</sup> and hip circumference adjusted for BMI<sup>27</sup>.

Once G×E loci have been identified, StructLMM can be used for the interpretation of these effects, and in particular to estimate per-individual allelic effects based on environmental profiles to identify individuals with increased or decreased trait risk (Fig. 4a). We confirmed the robustness of these estimates using hold-out validation, providing further evidence for possible opposite directions of effect at *PARK2* (Supplementary Fig. 17). To explore which environmental variables are most relevant for individual G×E signals, we calculated Bayes factors (BF) between the full model and models with individual environmental exposures removed (Supplementary Fig. 18), identifying between 20 and 25 environments with putative G×E effects (Bayes factors > 0). Because the environments are not independent of one another, we used backward elimination based on Bayes factors between the full model and models with increasing numbers of environments removed. These analyses identified physical activity measures for females (no evidence for males) as contributing to G×E at *MC4R*, in agreement with findings in ref. <sup>12</sup>, but also yielded a number of additional environments (Fig. 4b and Supplementary Fig. 18). For all loci, we consistently observed that multiple environments contribute to G×E, but there is evidence of differences in the G×E architecture, with *FTO* being associated with the largest number of environments and *SEC16B* and *PARK2* being associated with a smaller number of environments (Supplementary Fig. 18). Differences in the environments that contribute to G×E effects were also apparent when correlating per-individual allelic effect size estimates across loci (Supplementary Fig. 19).

**Identification of eQTL interactions with cellular state.** As a second application, we considered a gene expression dataset<sup>28</sup> to





**Fig. 3 | Applications to model GxE on BMI in UK Biobank.** **a**, Scatter plot of negative log  $P$  values from GxE interaction tests at 97 GIANT variants<sup>20</sup>, considering single-environment fixed-effect GxE tests (SingleEnv-Renv-int, x axis,  $P$  values Bonferroni adjusted for the number of tested environments) versus the StructLMM interaction test (StructLMM-int, y axis). Dashed lines correspond to  $\alpha < 0.05$ , Bonferroni adjusted for the number of tests. **b**, Local Manhattan plots of an interaction identified by StructLMM-int at *MC4R*. Top: LMM association test (LMM-Renv); middle, StructLMM interaction test (StructLMM-int); bottom, single-environment LMM interaction test (SingleEnv-Renv-int), for the environment with the strongest GxE effect at the GIANT SNP (vigorous physical activity x age). The red vertical line and diamond symbol indicates the GIANT SNP as in **a**. **c**, Scatter plot of genome-wide negative log  $P$  values from LMM association test (LMM-Renv, x axis) versus the StructLMM association test (y axis). Dashed lines indicate genome-wide significance at  $P < 5 \times 10^{-8}$ , and color denotes the estimated extent of heterogeneity (fitted parameter  $\rho$ ), where yellow and red correspond to variants with low and high GxE components, respectively. The inset shows a zoomed-in view of variants close to genome-wide significance.  $n = 252,188$  unrelated individuals of European ancestry for all experiments.

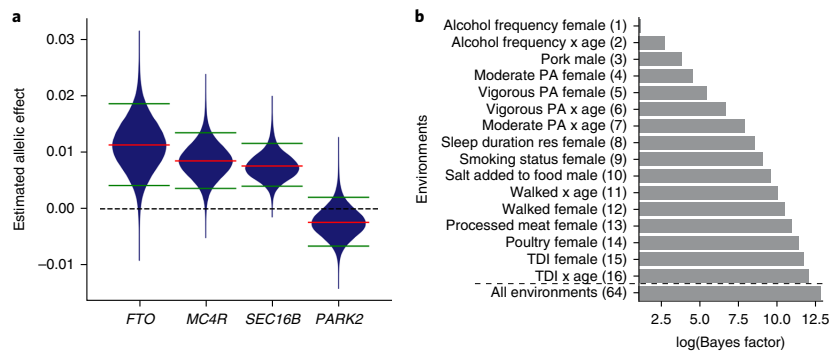
illustrate how StructLMM can be used to identify context-dependent regulatory effects on gene expression, for example due to external stimuli<sup>4</sup> or differences in cell type composition<sup>29</sup>, using hundreds of environment covariates. Insights into context-dependent genetic regulation of gene expression are important to identify disease-relevant cell types and molecular pathways<sup>30–32</sup>.

We reanalyzed a large whole-blood expression dataset comprising 2,040 genotyped individuals profiled with RNA-seq<sup>28</sup> (Methods) and applied StructLMM-int to test for cell-context interactions at *cis* expression quantitative trait loci (eQTL). Following Zhernakova et al.<sup>28</sup>, we considered gene expression levels both as phenotypes and as proxy (environmental) variables, which can tag variation in blood cell composition and other factors across individuals. Specifically, we considered a set of 443 highly variable genes as environmental variables in our analysis (Methods).

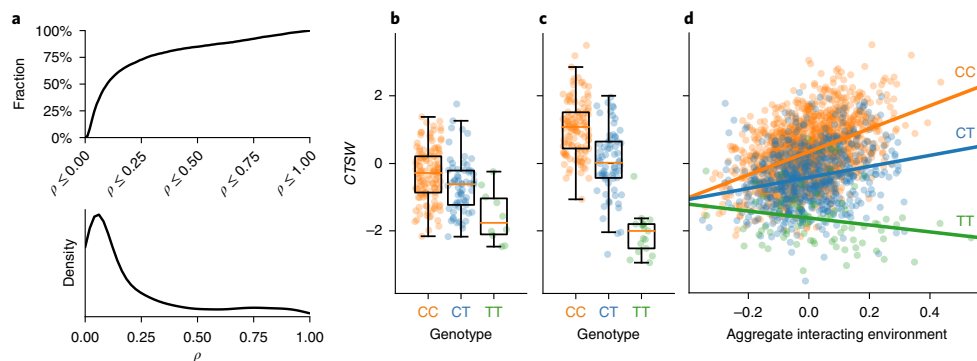
Initially, we applied a linear model to identify lead *cis* eQTL variants for 23,506 expressed genes (within  $\pm 250$  kb from the center of the gene, Methods). Next, we applied StructLMM-int to test for cell-context interactions at lead variants for each of these genes. The model produced calibrated  $P$  values despite the large number of environments (Supplementary Fig. 20), identifying 3,483 eQTL with a cell-context interaction (FDR  $< 5\%$ , termed interaction eQTL; Supplementary Table 5). Although globally interactions with cell context tended to explain small fractions of the *cis* genetic variance

on gene expression ( $\rho < 0.2$ , for 68.0% of interaction eQTL; Fig. 5a), GxE explained more variance than persistent genetic effects for 532 genes ( $\rho > 0.5$ , for 15.3% of interaction eQTL). We also compared StructLMM-int to alternative multi-environment interaction tests based on fixed effects, which were markedly less robust and identified fewer interaction eQTL (Supplementary Fig. 20). Similarly, we compared the discovered interaction eQTL to results from a stepwise procedure that was used to identify interaction eQTL in the primary analysis of the same data<sup>28</sup> (details in Supplementary Note), which yielded markedly fewer interactions (3,372 versus 1,841 interaction eQTL, considering StructLMM and the approach in ref. 28; FDR  $< 5\%$ ; Supplementary Fig. 20; considering 17,952 genes assessed in both studies). Finally, we considered alternative approaches to normalize the expression data (Methods), thereby assessing potential biases due to gene-exposure correlations and distributional skew of counts-based gene expression profiles. These results indicated that StructLMM is robust to both potential sources of bias (Supplementary Fig. 21).

Next, we overlapped the interaction eQTL with risk variants from the NHGRI-EBI GWAS catalog V1.0.1 (ref. 33), identifying 64 putative colocalization events ( $r^2 > 0.8$  between lead eQTL and GWAS variants; Supplementary Fig. 22 and Methods), including GWAS variants for autoimmune diseases, infectious diseases and blood cell traits (Supplementary Table 6 and Supplementary Datasets 1 and 2).



**Fig. 4 | Downstream analysis to explore identified GxE loci. a**, Violin plots showing distributions of the in-sample estimated allelic effect size (effect of heterozygous versus homozygous reference carriers for environmental states realized in the population;  $n = 252,188$  unrelated individuals of European ancestry for all experiments; Methods) on BMI for the four GIAN variants with GxE ( $\alpha < 0.05$ , Fig. 3a). Estimated persistent genetic effects are shown by the red bar, and the green bars indicate top and bottom 5% quantiles of variation in effect sizes due to GxE. **b**, Cumulative evidence of environmental variables (PA, physical activity; TDI, Townsend deprivation index) that explain GxE at MC4R, showing Bayes factors between the full model and models with increasing numbers of environmental variables removed using backward elimination. For comparison, evidence for all 64 environmental variables is shown. 'Alcohol frequency female' is selected as the first environmental factor, followed by 'Alcohol frequency x age' and so on.



**Fig. 5 | Gene-context interactions in a blood gene expression cohort. a**, Cumulative fraction (top) and density (bottom) of eQTL with interactions (3,483 interaction eQTL; FDR  $< 5\%$ ) as a function of the estimated extent of heterogeneity (fitted parameter  $\rho$ ). **b–d**, Example of an interaction eQTL for CTSW at the lead variant rs568617, which is in LD with rs568617 ( $r^2 = 0.98$ , Supplementary Fig. 23), a known risk variant for Crohn's disease. **b, c**, Expression level of CTSW for different alleles at the lead eQTL variant, considering 10% strata of individuals ( $n = 204$  independent samples) with the smallest (**b**) and largest (**c**) per-individual allelic effects, as estimated using StructLMM, displaying the 25<sup>th</sup>, 50<sup>th</sup> and 75<sup>th</sup> percentiles, with whiskers extending to 1.5 $\times$  the interquartile range. **d**, Scatter plot of CTSW expression level versus the aggregate environmental signal for the GxE effect at rs568617 (aggregate interacting environment), estimated using StructLMM (Supplementary Note). Individuals are stratified by the alleles at the eQTL lead variant. Solid lines denote regression lines for each genotype group.

Notably, 46 of these interaction eQTL were not reported in the primary analysis<sup>38</sup>. One example is an interaction eQTL for CTSW expression (Fig. 5b–d,  $P$  StructLMM-int =  $2.2 \times 10^{-15}$ ,  $\rho = 0.12$ ), which is in linkage disequilibrium (LD) with a risk variant for Crohn's disease rs568617 ( $r^2 = 0.98$ , Supplementary Fig. 23). To investigate the molecular pathways that are associated with this interaction, we stratified the population into strata with the smallest and largest allelic effects, as estimated using StructLMM (Fig. 5b,c), and tested for pathways that were enriched among differentially expressed genes between these groups (Methods). This identified *T cell selection* (GO: 0045058), *positive T cell selection* (GO: 0046632) and *positive regulation of interleukin-17 secretion* (GO: 0032740) as the top three processes for this interaction eQTL (Fisher exact test; genome-wide enrichment results in Supplementary Table 6), GO terms that are consistent with known

roles of IL-17-producing CD4<sup>+</sup> T cells in the pathogenesis of inflammatory bowel disease, including Crohn's disease<sup>34</sup>.

Taken together, results from this analysis demonstrate the broad applicability of StructLMM, including in settings with large numbers of environmental factors.

## Discussion

We propose a method based on variance component tests to identify GxE interactions using multiple environments. Conceptually, our approach is related to set tests for groups of variants, but instead of aggregating across multiple genetic variants, StructLMM jointly models multiple environmental variables to identify GxE interactions. Compared with conventional single and multiple degrees of freedom fixed-effect GxE tests, this approach enjoys

power advantages (Fig. 2 and Supplementary Figs. 1–4) and yields increased robustness, in particular when analyzing large numbers of environmental variables (Supplementary Figs. 1 and 20).

We applied StructLMM to data from UK Biobank to assess G×E at 97 GIANT variants associated with BMI, confirming established G×E effects at *FTO*, and we identified, for the first time, three additional G×E signals at stringent thresholds (Family Wise Error Rate (FWER) < 5%; Fig. 3a), some of which confirm prior evidence<sup>11,12,14,15,22,23</sup>. More-lenient FDR-based significance thresholds, as frequently employed for G×E analyses<sup>5,12</sup>, yielded 11 GIANT variants with evidence for G×E (FDR < 5%; Benjamini-Hochberg adjusted; Supplementary Table 3), and a genome-wide analysis based on all variants that are associated with BMI identified 23 loci with significant G×E effects (Supplementary Fig. 11 and Supplementary Table 3). We also show that the same framework can be used to test for associations, demonstrating that accounting for heterogeneity in effect sizes can identify additional loci, similarly to previously reported benefits of two degrees of freedom fixed-effect tests<sup>5</sup>.

In addition to offering power advantages, StructLMM yields per-individual allelic effect size estimates that reflect G×E. We have shown that this allows for different downstream analyses, including the identification of individuals with increased or decreased genetic risk. This would be of particular interest in the complex disease field, as it may provide further explanation as to why individuals who share the same set of risk variants may have different outcomes in longitudinal follow-up. In particular, identifying sets of environments that may decrease disease risk for individuals carrying the same genetic burden may provide useful avenues for targeted disease prevention. We also explore which environments are putative drivers of the observed G×E effects. However, such downstream analyses, when using the same dataset for discovery, should be interpreted with caution. Ultimately, independent validation cohorts will be required to confirm such findings.

As a second-use case, we applied StructLMM to test for cell-context interactions in a large blood eQTL study, where the same modeling principles enabled the identification of context-specific eQTL. Several of these interaction eQTL colocalized with GWAS variants, and the marker genes of the cellular environments that underlie these interaction effects could be connected to plausible molecular pathways (Supplementary Table 6).

Although we found that StructLMM is a robust and powerful alternative to conventional linear interaction tests, our approach is not free of limitations. First, there are general challenges when analyzing G×E that although not specific to our model need to be taken into consideration. One such challenge is environmental variables that are themselves heritable. Accounting for heritable covariates in association tests can lead to spurious associations due to collider bias<sup>35</sup>. Our results indicate that interaction tests are more robust to such correlations (Supplementary Fig. 3). However, gene exposure associations alter the interpretation of interactions, reflecting epistatic relationships between genetic factors. A second generic challenge is the selection of candidate variants for G×E tests. To reduce the multiple testing burden, we selected variants that have persistent effects on the phenotype. However, the fact that our association test identifies novel loci with strong G×E ( $\rho$ ) if applied genome-wide indicates that this filter is not optimal.

Among more specific limitations and areas of future work for StructLMM, we note the computational requirements of the model are more demanding than conventional LMMs, despite scaling linearly with the number of individuals. A second potential limitation is that StructLMM does not currently enable accounting for relatedness. Although the model has an additive random effect component, it is currently used to model additive environmental effects. Generalizations to simultaneously account for a relatedness could be considered, for example, using suitable low-rank approximations<sup>36</sup> or other speed-ups to retain scalability to large sample

sizes. Finally, although StructLMM can in principle be used in conjunction with any environmental covariance, we have limited our attention to linear covariances. The model could be extended to account for non-linear interactions, for example, using polynomial covariance functions. Future developments in this direction will be increasingly valuable as larger cohort sizes enable detection of higher-order interaction effects.

**URLs.** Haplotype Reference panel, <http://www.haplotype-reference-consortium.org/site>; Phase 3 1000 Genomes reference panel, <http://grch37.rest.ensembl.org>.

#### Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41588-018-0271-0>.

Received: 6 March 2018; Accepted: 4 October 2018;  
Published online: 26 November 2018

#### References

- Hunter, D. J. Gene-environment interactions in human diseases. *Nat. Rev. Genet.* **6**, 287–298 (2005).
- Ritz, B. R. et al. Lessons learned from past gene-environment interaction successes. *Am. J. Epidemiol.* **186**, 778–786 (2017).
- Brown, A. A. et al. Genetic interactions affecting human gene expression identified by variance association mapping. *eLife* **3**, e01381 (2014).
- Fairfax, B. P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
- Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J. & Gauderman, W. J. Exploiting gene-environment interaction to detect genetic associations. *Hum. Hered.* **63**, 111–119 (2007).
- Rask-Andersen, M., Karlsson, T., Ek, W. E. & Johansson, A. Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. *PLoS Genet.* **13**, e1006977 (2017).
- Lin, X., Lee, S., Christiani, D. C. & Lin, X. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* **14**, 667–681 (2013).
- Lin, X. et al. Test for rare variants by environment interactions in sequencing association studies. *Biometrics* **72**, 156–164 (2016).
- Casale, F. P., Horta, D., Rakitsch, B. & Stegle, O. Joint genetic analysis using variant sets reveals polygenic gene-context interactions. *PLoS Genet.* **13**, e1006693 (2017).
- Kilpelainen, T. O. et al. Physical activity attenuates the influence of *FTO* variants on obesity risk: a meta-analysis of 218,166 adults and 19,268 children. *PLoS Med.* **8**, e1001116 (2011).
- Ahmad, S. et al. Gene x physical activity interactions in obesity: combined analysis of 111,421 individuals of European ancestry. *PLoS Genet.* **9**, e1003607 (2013).
- Bjornland, T., Langaas, M., Grill, V. & Mostad, I. L. Assessing gene-environment interaction effects of *FTO*, *MC4R* and lifestyle factors on obesity using an extreme phenotype sampling design: Results from the HUNT study. *PLoS One* **12**, e0175071 (2017).
- Young, A. L., Wauthier, F. & Donnelly, P. Multiple novel gene-by-environment interactions modify the effect of *FTO* variants on body mass index. *Nat. Commun.* **7**, 12724 (2016).
- Corella, D. et al. Statistical and biological gene-lifestyle interactions of *MC4R* and *FTO* with diet and physical activity on obesity: new effects on alcohol consumption. *PLoS One* **7**, e52344 (2012).
- Qi, Q. et al. Fried food consumption, genetic risk, and body mass index: gene-diet interaction analysis in three US cohort studies. *BMJ* **348**, g1610 (2014).
- Lee, S. et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–237 (2012).
- Crawford, L., Zeng, P., Mukherjee, S. & Zhou, X. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLoS Genet.* **13**, e1006869 (2017).
- Gauderman, W. J. et al. Update on the state of the science for analytical methods for gene-environment interactions. *Am. J. Epidemiol.* **186**, 762–770 (2017).
- The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

20. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
21. Bycroft, C. et al. Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*, <https://doi.org/10.1101/166298> (2017).
22. Richardson, A. S. et al. Moderate to vigorous physical activity interactions with genetic variants and body mass index in a large US ethnically diverse cohort. *Pediatr. Obes.* **9**, e35–e46 (2014).
23. Ahmad, S. et al. Established BMI-associated genetic variants and their prospective associations with BMI and other cardiometabolic traits: the GLACIER Study. *Int. J. Obes. (Lond.)* **40**, 1346–1352 (2016).
24. Hall, N. G., Klenotic, P., Anand-Apte, B. & Apte, S. S. ADAMTSL-3/punctin-2, a novel glycoprotein in extracellular matrix related to the ADAMTS family of metalloproteases. *Matrix Biol.* **22**, 501–510 (2003).
25. Zillikens, M. C. et al. Large meta-analysis of genome-wide association studies identifies five loci for lean body mass. *Nat. Commun.* **8**, 80 (2017).
26. Wen, W. et al. Genome-wide association studies in East Asians identify new loci for waist-hip ratio and waist circumference. *Sci. Rep.* **6**, 17958 (2016).
27. Shungin, D. et al. New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).
28. Zhernakova, D. V. et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
29. Westra, H. J. et al. Cell specific eQTL analysis without sorting cells. *PLoS Genet.* **11**, e1005223 (2015).
30. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10**, 184–194 (2009).
31. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
32. Emilsson, V. et al. Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).
33. MacArthur, J. et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).
34. Galvez, J. Role of Th17 cells in the pathogenesis of human IBD. *ISRN Inflamm.* **2014**, 928461 (2014).
35. Day, F. R., Loh, P.-R., Scott, R. A., Ong, K. K. & Perry, J. R. A robust example of collider bias in a genetic association study. *Am. J. Hum. Genet.* **98**, 392–393 (2016).
36. Listgarten, J., Lippert, C. & Heckerman, D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nat. Genet.* **45**, 470 (2013).

### Acknowledgements

The authors thank C. Lippert and L. Parts for helpful discussions. This research was conducted using the UK Biobank Resource (Application Number 14069). R.M. was supported by a PhD fellowship from the Mathematical Genomics and Medicine program, funded by the Wellcome Trust. F.P.C., D.H. and O.S. received support from core funding of the European Molecular Biology Laboratory and the European Union's Horizon2020 research and innovation program under grant agreement N635290. I.B. acknowledges funding from Wellcome (WT098051 and WT206194). M.J.B. was supported by a fellowship from the EMBL Interdisciplinary Postdoc (EI3POD) program under Marie Skłodowska-Curie Actions COFUND (grant number 664726). The Biobank-Based Integrative Omics Studies (BIOS) Consortium is funded by BBMRI-NL, a research infrastructure financed by the Dutch government (NWO 184.021.007).

### Author contributions

R.M., F.P.C., I.B. and O.S. conceived the method. R.M., F.P.C., and D.H. implemented the methods. R.M., F.P.C., and M.J.B. analyzed the data. L.F. and BIOS Consortium provided data resources. R.M., F.P.C., I.B., and O.S. interpreted results and wrote the paper.

### Competing interests

F.P.C. was employed at Microsoft while performing the research.

### Additional information

**Supplementary information** is available for this paper at <https://doi.org/10.1038/s41588-018-0271-0>.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Correspondence and requests for materials** should be addressed to I.B. or O.S.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2018

## BIOS Consortium

**Bastiaan T. Heijmans<sup>10</sup>, Peter A. C.'t Hoen<sup>11</sup>, Joyce van Meurs<sup>12</sup>, Aaron Isaacs<sup>13</sup>, Rick Jansen<sup>14</sup>, Lude Franke<sup>15</sup>, Dorret I. Boomsma<sup>16</sup>, René Pool<sup>16</sup>, Jenny van Dongen<sup>16</sup>, Jouke J. Hottenga<sup>16</sup>, Marleen M. J. van Greevenbroek<sup>17</sup>, Coen D. A. Stehouwer<sup>17</sup>, Carla J. H. van der Kallen<sup>17</sup>, Casper G. Schalkwijk<sup>17</sup>, Cisca Wijmenga<sup>15</sup>, Alexandra Zhernakova<sup>15</sup>, Ettje F. Tigchelaar<sup>15</sup>, P. Eline Slagboom<sup>10</sup>, Marian Beekman<sup>10</sup>, Joris Deelen<sup>10</sup>, Diana van Heemst<sup>18</sup>, Jan H. Veldink<sup>13</sup>, Leonard H. van den Berg<sup>13</sup>, Cornelia M. van Duijn<sup>19</sup>, Bert A. Hofman<sup>20</sup>, André G. Uitterlinden<sup>12</sup>, P. Mila Jhamai<sup>12</sup>, Michael Verbiest<sup>11</sup>, H. Eka D. Suchiman<sup>10</sup>, Marijn Verkerk<sup>12</sup>, Ruud van der Breggen<sup>10</sup>, Jeroen van Rooij<sup>12</sup>, Nico Lakenberg<sup>10</sup>, Hailiang Mei<sup>21</sup>, Maarten van Iterson<sup>10</sup>, Michiel van Galen<sup>11</sup>, Jan Bot<sup>22</sup>, Peter van't Hof<sup>21</sup>, Patrick Deelen<sup>15</sup>, Irene Nooren<sup>22</sup>, Matthijs Moed<sup>10</sup>, Martijn Vermaat<sup>11</sup>, Dasha V. Zhernakova<sup>15</sup>, René Luijk<sup>10</sup>, Marc Jan Bonder<sup>15</sup>, Freerk van Dijk<sup>15,23</sup>, Wibowo Arindrarto<sup>24</sup>, Szymon M. Kielbasa<sup>21</sup>, Morris A. Swertz<sup>15,23</sup> and Erik W. van Zwet<sup>24</sup>**

<sup>10</sup>Molecular Epidemiology Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands.

<sup>11</sup>Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands. <sup>12</sup>Department of Internal Medicine, ErasmusMC, Rotterdam, the Netherlands. <sup>13</sup>Department of Neurology, Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht, the Netherlands. <sup>14</sup>Department of Psychiatry, VU University Medical Center, Neuroscience Campus Amsterdam, Amsterdam, the Netherlands. <sup>15</sup>Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, the Netherlands. <sup>16</sup>Department of Biological Psychology, VU University Amsterdam, Neuroscience Campus Amsterdam, Amsterdam, the Netherlands. <sup>17</sup>Department of Internal Medicine and School for Cardiovascular Diseases (CARIM), Maastricht University Medical Center, Maastricht, the Netherlands. <sup>18</sup>Department of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, the Netherlands. <sup>19</sup>Department of Genetic Epidemiology, ErasmusMC, Rotterdam, the Netherlands. <sup>20</sup>Department of Epidemiology, ErasmusMC, Rotterdam, the Netherlands. <sup>21</sup>Sequence Analysis Support Core, Leiden University Medical Center, Leiden, the Netherlands. <sup>22</sup>SURFsara, Amsterdam, the Netherlands. <sup>23</sup>Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands. <sup>24</sup>Medical Statistics Section, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, the Netherlands.

## Methods

**The structured linear model.** A conventional LMM to test for associations can be cast as:

$$y = Xb + x\beta + u + \psi$$

where  $\beta$  is the focal variant effect size,  $X$  is the fixed-effect design matrix of  $K$  covariates, and  $b$  is the corresponding effect size. The variable  $u$  denotes additive (confounding) factors, and  $\psi$  denotes i.i.d. noise. The random effect component  $u$  and the noise vector  $\psi$  follow multivariate normal distributions,  $u \sim N(0, \sigma_u^2 \Sigma_u)$  and  $\psi \sim N(0, \sigma_\psi^2 I)$ , where the covariance matrix  $\Sigma_u$  reflects the covariance of population structure, environment or other (confounding) factors. Association tests for non-zero effects of the focal variant correspond to alternative hypothesis  $\beta \neq 0$ .

StructLMM generalizes the conventional LMM for association testing by introducing per-individual effect sizes due to G×E

$$y = Xb + x\beta + x \odot \beta_{G \times E} + u + \psi \quad (2)$$

where  $\beta_{G \times E}$  is a per-individual allelic effects vector that follows a multivariate normal distribution with environment covariance  $\Sigma$ :

$$\beta_{G \times E} \sim N(0, \sigma_{G \times E}^2 \Sigma) \quad (3)$$

The covariance  $\Sigma$  captures heterogeneity in allelic effects in the population and is estimated using a linear covariance function based on a set of observed environmental variables, where we assume  $\Sigma_u = \Sigma$ . If collider bias<sup>35</sup> is a concern, non-heritable environmental variables should be selected. Non-linear environmental effects can be modeled by combining observed environmental variables (for example, effects from environments × age or environments × gender; Supplementary Note).

**Statistical testing.** Based on equation (1), we define an interaction test ( $\sigma_{G \times E}^2 > 0$ ) where persistent genetic and additive environment effects are accounted for in the null model and an association test ( $\sigma_{G \times E}^2 > 0$  and  $\beta \neq 0$ ), which jointly tests for associations while accounting for the possibility of heterogeneous genetic effects due to G×E. Both tests are implemented as efficient score tests, similar to the approach in SKAT and SKAT-O<sup>37,38</sup>, with linear complexity in the number of individuals (Supplementary Note).

**Estimation of  $\rho$ .** Estimates of the fraction of the genetic variance explained by G×E ( $\rho$ ) can be obtained from maximum likelihood estimates of the model in equation (4).

$$\rho = \frac{\text{Var}^{G \times E}}{\text{Var}^G + \text{Var}^{G \times E}} \quad (4)$$

with  $\text{Var}^G$  denoting the fraction of the variance explained by persistent effects and  $\text{Var}^{G \times E}$  denoting variance due to G×E.

**Exploring the most relevant environments for G×E.** Bayes factors between the full model and models with individual environments or sets of environments removed from the environmental covariance  $\Sigma$  (Supplementary Note) can be used to assess the relevance of environments.

**Estimation of per-individual allelic effects.** Per-individual (for each environment state) allelic effects can be estimated using BLUP<sup>39</sup>.

Additionally, the model yields posterior estimates of the realization of the unobserved environmental state that explains the G×E effect (Supplementary Note).

**Simulations.** *Simulation procedure overview.* Simulations were based on genotypes of European individuals from the 1000 Genomes project<sup>19</sup> (phase 1, 1,092 individuals, 379 Europeans), considering 103,527 variants on chromosome 21 (minor allele frequency  $\geq 2\%$ ). Following refs<sup>40,41</sup>, we generated synthetic genotypes of unrelated individuals for different sample sizes while preserving the population structure of the seed population (as in ref.<sup>9</sup>). We considered 33 environmental exposures using empirical environmental covariates from 70,282 UK Biobank individuals (based on the Interim release), augmented using element-wise interactions with gender and age, resulting in 100 environmental variables. These environmental variables were preprocessed as in the UK Biobank analysis (discussed below) and randomly assigned to synthetic genotypes (details in Supplementary Note).

*Assessment of statistical calibration.* Statistical calibration of different tests was assessed using phenotypes simulated from an empirical null model, considering (i) no genetic effect (Fig. 2a and Supplementary Fig. 1a,b) and (ii) simulated persistent genetic effects (100 persistent genetic effect variants, no G×E interactions; Supplementary Fig. 1a,b). Calibration was assessed using QQ plots and genomic

control ( $\lambda_{GC} = \frac{\log_{10}(m)}{\log_{10}(0.5)}$ ;  $m$  is the median  $P$  value), based on  $P$  values from chromosome 21 pooled across 100 repeat experiments.

*Power simulations.* Phenotypes with G×E interactions were simulated, varying the fraction of variance explained by G×E, the number of active environments and other parameters (Supplementary Table 1 and Supplementary Note). We also studied the effect of gene-exposure correlations (Supplementary Fig. 3 and Supplementary Note) and considered synthetic environments to assess the effect of (rare) binary environmental variables (Supplementary Fig. 4 and Supplementary Note). We considered 1,000 repeat experiments for each setting, randomly selecting a segment of approximately 2 Mb from chromosome 21 and simulating G×E effects from one causal variant. Power at 1% FWER (Bonferroni adjusted across variants) was assessed considering variants in linkage disequilibrium with selected true causal variants ( $r^2 \geq 0.8$ ) as true positives, reporting average power across repeat experiments (individual experiments return 1 or 0).

*Comparison methods.* We compared StructLMM to alternative single- and multi-environment models, as well as standard genetic association tests. For interaction tests, we considered alternative single-environment G×E interaction tests (i) using random effect (SingleEnv-Renv-int) or (ii) fixed-effect (SingleEnv-Fenv-int) components to account for additive environmental effects due to all environments, and finally (iii) an additive single-environment fixed effect term based on the specific environment considered in the G×E test only (SingleEnv-Senv-int). The same models were considered to test for associations, using a two degrees of freedom (df) statistical test<sup>4</sup> (SingleEnv-Renv, SingleEnv-Fenv, SingleEnv-Senv, respectively). Additionally, for association tests, we considered linear association tests, again either using a multi-environment random effect for additive environment (LMM-Renv) or a multi-environment fixed effect (LM-Fenv) component to account for additive environmental effects, as well as a linear model with no additive environment effect term (LM). All tests were implemented using LRT, considering Bonferroni-adjusted minimum  $P$  value per variant across environments for single-environment models. Finally, we assessed the performance of fixed-effect multi-environment interaction and association tests, again considering either random or fixed additive environment components based on all observed environments, considering either an LRT or score test. Performance of these multi-environment tests was assessed using the average area under the curve (AUC) across repeat experiments (using true positive definitions as for power), computed in the range FPR  $< 0.10$  and normalized to the 0–1 range, such that 0 corresponds to chance performance and 1 is the performance of an ideal model (Supplementary Fig. 1c,d). An overview of all methods compared is provided in Supplementary Table 2 and details in Supplementary Note.

**Analysis of BMI in UK Biobank.** This research was conducted using the full release of the UK Biobank Resource (Application 14069)<sup>31</sup>. The UK Biobank study has approval from the North West Multi-Centre Research Ethics Committee and all participants included in the analyses provided informed consent to UK Biobank.

*Data pre-processing.* BMI phenotype data are 'Instance 0' of UK Biobank data field 21001. Individuals with missing BMI data were discarded from the analysis and BMI log transformed<sup>33,42</sup>. Following ref.<sup>15</sup>, we considered 21 lifestyle covariates as environments, discarding individuals with outlying or missing environmental variables (Supplementary Note). We further discarded individuals of non-British ancestry and related individuals. After filtering and QC on the BMI phenotype, genotype and the environmental variables, we obtained a set of 252,188 individuals for analysis. Principal components for population structure adjustment were calculated using flashpca version 2.0 (ref.<sup>43</sup>) using 147,604 variants, as indicated by the field 'in\_PCA' from the released marker QC file.

*Genotype data.* We used genotypes that were imputed with the HRC panel (build GRCh37). We performed QC of the imputed variants on the fly, using a fast bgen reader, implemented as part of StructLMM, treating genotype-sample pairs with low imputation accuracy (max. probability  $< 0.5$ ) as missing and discarding variants with missingness  $> 5\%$ , MAF  $< 1\%$ , HWE  $P < 1 \times 10^{-6}$  and INFO score  $r^2 < 0.4$  (based on the UK Biobank imputation MAF and info file). Genotype dosages of remaining variants were calculated using available probabilities (including genotype-sample pairs with low imputation accuracy) and mean imputation used for any genotype-sample pairs with missing data. 7,515,856 variants passed these filters.

*Environmental covariance and covariates.* To generate the environment matrix  $E$ , we augmented all 21 environmental variables described above (excluding age) by gender and age by multiplying the continuous age vector, the binary male indicator vector and the binary female indicator vector with each of the 21 environment variables, which resulted in 63 covariates. The environmental covariance was estimated based on standardized environmental variables (not including zero values due to augmentation when mean adjusting) followed by per-individual standardization (Supplementary Figs. 5 and 6; full details in Supplementary Note). In all analyses,



a mean vector, genotype chip, gender, age<sup>2</sup>, age<sup>3</sup>, gender × age, gender × age<sup>2</sup>, gender × age<sup>3</sup>, 10 genetic principal components were included as covariates.

**Calibration of interaction and association tests.** To validate the tested methods and QC procedures, we assessed the empirical calibration using permuted genotype variants (173,297 variants) on chromosome 20 (Supplementary Fig. 7).

**Interaction testing.** We considered 97 GIANT variants previously associated with BMI<sup>20</sup> to test for G×E interactions using StructLMM-int and single-environment fixed-effect interaction tests (SingleEnv-Renv-int and SingleEnv-Senv-int, 1 df, Supplementary Note) and a multi-environment fixed-effect-based interaction test (64 df, MultiEnv-Renv-LRT-int, Supplementary Note). Variants with significant G×E were reported at FWER 5% ( $P < 0.05/97$ ), and alternatively using a more lenient threshold at FDR < 5% (Benjamini–Hochberg adjustment<sup>41</sup>; Supplementary Table 3).

We also selected the 17,606 variants with LMM-Renv  $P$  values <  $5 \times 10^{-8}$  and compared results using StructLMM-int to those from single-environment fixed-effect interaction tests (SingleEnv-Renv-int, Supplementary Note) and a multi-environment fixed-effect-based interaction test (64 df, MultiEnv-Renv-LRT-int, Supplementary Note). This filter is valid, as LMM-Renv corresponds to the null model of both StructLMM-int and SingleEnv-Renv-int. Variants with significant G×E were reported at FDR < 5% (Benjamini–Hochberg<sup>41</sup> adjustment), followed by LD clumping to define independent loci: we iteratively (i) selected the most significant variant (using the FDR-adjusted  $P$  values) and (ii) removed all variants in LD ( $r^2 > 0.1$ ) within ±500 kb, until no variant was left, resulting in 23, 11 and 9 clumps (loci), respectively (Supplementary Table 3).

**Association testing.** We used StructLMM, LMM-Renv and LM for genome-wide association analyses, reporting significant associations at  $P < 5 \times 10^{-8}$ , for which  $\rho$  was estimated using StructLMM (Fig. 3c and Supplementary Table 4). LD clumping was used to define independent loci identified by each of the three methods; we iteratively (i) selected the most significant variant and (ii) removed all variants in LD ( $r^2 > 0.1$ ) within ±500 kb, until no variant was left, resulting in 351, 327 and 379 loci, respectively. We compared the methods pairwise, identifying loci found by only one method, by calculating the LD ( $r^2$ ) between significant variants within a clump identified by one method and significant variants from the other method that lie within ±500 kb, resulting in 32 and 16 loci (StructLMM and LMM-Renv), 65 and 98 loci (StructLMM and LM) and 47 and 97 loci (LMM-Renv and LM). We also compared the genome-wide results of StructLMM to those from a multiple degrees of freedom (65 df) fixed-effect association test MultiEnv-Renv-LRT, again using genome-wide significance thresholds of  $5 \times 10^{-8}$  and estimating  $\rho$  for all significant variants (Supplementary Fig. 9).

**Per-individual allelic effect estimation.** We performed in-sample estimation of the allelic effect (discussed above) for each of 252,188 individuals at each of the four interaction loci (FWER 5%; Fig. 4a). Allelic effects were assessed out of sample by randomly splitting the cohort into training and test fractions, to assess out-of-sample predictions (Supplementary Fig. 17 and Supplementary Note). To assess whether the same set of individuals are at the extreme ends of the effect size spectrum across multiple interaction variants (5% FDR-adjusted), we computed the squared Spearman's correlation coefficient and then used Ward hierarchical clustering (Supplementary Fig. 19).

**Explorative analysis of driving environments.** We explored which environments had putative effects on G×E by comparing the log marginal likelihood of the full model to models with individual or sets of environments excluded. We initially assessed the relevance of individual environments based on the log(Bayes factor) of removing single environments (Supplementary Fig. 18). To account for correlations between environments, we also used a backwards elimination procedure (Supplementary Note), greedily removing environments until there was evidence that we selected a full set of environments that can drive the observed G×E effect (Fig. 4b and Supplementary Fig. 18).

**Analysis of cell-context eQTL in a large blood cohort.** *Genotype data pre-processing.* We used freeze one from the BIOS consortium (EGA; accession EGAS00001001077) and analyzed 2,040 samples for which genotypes and QC-passing RNA-seq data were available. Processed genotype and expression data were taken from the primary analysis<sup>20</sup>. Imputed genotypes (from the four biobanks CODAM, LifeLines, the Leiden Longevity Study and the Rotterdam Study) were merged to perform a mega-analysis, as opposed to the meta-analysis in the original paper. After merging, we performed joint QC of the genetic variants, retaining variants that met the following conditions: MACH-R2 > 0.5, call rate > 0.95, HWE >  $1 \times 10^{-4}$  and MAF > 5%, resulting in 5,683,643 variants for analysis.

**Ethical approval.** The ethical approval for this study lies with the individual participating cohorts (CODAM, LLD, LLS and RS)<sup>45–48</sup>.

**Expression data.** The expression data were taken from the original quantifications (after TMM normalization), and we selected features that were identified in at least

10% of the samples, resulting in 23,506 expressed genes for analysis. Expression values were quantile normalized, and we used ENSEMBL 71 as gene annotation.

**Environmental covariance.** We used gene expression levels to build the StructLMM covariance, capturing cell-type composition and other sources of cell-context heterogeneity. Specifically, we considered a set of highly variable proxy genes, identified through a two-step procedure: (i) we selected the top 25% most variable genes based on the interquartile range of non-quantile normalized data, (ii) we pruned this set, ranking the genes by variability and removing genes with  $r^2 \geq 0.2$  with a higher-ranked feature. This method resulted in a set of 443 proxy genes, which we used to build a linear covariance for StructLMM based on quantile-normalized expression levels.

**cis-eQTL map.** We identified cis-eQTL using a linear association test, considering genetic variants within 250 kb from the center of the gene body. After the primary analysis<sup>20</sup>, we considered the following 53 factors as covariates: the first 25 principal components calculated from the full gene expression profiles, the leading ten MDS components on the genotypes (computed using PLINK v1.90b3.32); cell counts of neutrophils, eosinophils, basophils, lymphocytes and monocytes; age; gender; dataset batch; and the first eight principal components derived from SAMtools flagstat and Picard tools (Supplementary Note).

**Interaction eQTL analysis.** For each of the 23,506 genes, we tested for interactions at the lead variant from the cis-eQTL map using StructLMM-int. For comparison, we also considered a multivariate fixed effect test, MultiEnv-Renv-LRT-int (Supplementary Note) either using the same environmental variables as in StructLMM or based on a reduced representation using the leading twenty principal components (Supplementary Note). Significant interactions were reported at FDR < 5% (Storey's procedure<sup>49</sup>). Calibration of all methods was assessed by repeating the analysis with permuted genotypes. We considered analogous analyses using residual gene expression levels as environments, regressing out the cis genetic variant tested from all environments (Supplementary Fig. 21a–c), to rule out potential spurious effects due to strong gene-exposure correlations. As an additional control, we considered an alternative normalization of the expression data using boxcox normalization followed by removal of outliers (2.5 standard deviations, Supplementary Fig. 21d–f).

**Overlap with GWAS hits and pathways analysis.** We overlapped our set of interaction eQTL with GWAS variants that are part of the NHGRI-EBI GWAS catalog<sup>51</sup> that pass the genome-wide significance threshold ( $P < 5 \times 10^{-8}$ ). We defined a colocalization event based on (i) eQTL and GWAS variants are within 10 kb and (ii) high linkage disequilibrium between variants ( $r^2 \geq 0.8$ , estimated from Phase 3 1000 Genomes reference panel). For the pathway enrichment analysis, we used the following procedure for each analyzed interaction eQTL: (i) we used StructLMM to predict per-individual allelic effects (described above); (ii) we defined the groups of samples with the highest or lowest predicted allelic effect, each containing 10% of the total number of samples ( $N = 204$ ); (iii) we computed rank-based correlation of genome-wide expression levels and the vector of group binary indicators (based on  $N = 408$  samples); (iv) we defined the 100 genes with the highest positive correlation as differentially expressed; (v) we performed enrichment analysis of GO biological processes in the differentially expressed test using topGO<sup>50</sup> (standard Fisher's exact test, algorithm = classic, nodeSize = 5). In Supplementary Table 6, we report both the top-enriched broad biological process and the three top-enriched narrow processes (broad/narrow terms are defined as those with more/less than 100 annotated genes in the background set). In the CTSW example in Fig. 5, the aggregate interacting environment was estimated as described above.

Further statistical details and derivatives are provided in Supplementary Note.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

**Code availability.** StructLMM is available from <https://github.com/limix/struct-lmm> and is supported within the LIMIX framework<sup>51</sup> at <https://github.com/limix/limix>. For tutorials and illustrations on how to use the model, see <http://struct-lmm.readthedocs.io>.

## Data availability

The BIOS RNA data can be obtained from the European Genome-phenome Archive (EGA; accession EGAS00001001077). Genotype data are available from the respective biobanks.

## References

- Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).
- Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).

39. Schaeffer, L. Application of random regression models in animal breeding. *Livest. Prod. Sci.* **86**, 35–45 (2004).
40. Casale, F. P., Rakitsch, B., Lippert, C. & Stegle, O. Efficient set tests for the genetic analysis of correlated traits. *Nat. Methods* **12**, 755–758 (2015).
41. Loh, P. R. et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
42. Fesinmeyer, M. D. et al. Genetic risk factors for BMI and obesity in an ethnically diverse population: results from the population architecture using genomics and epidemiology (PAGE) study. *Obesity* **21**, 835–846 (2013).
43. Abraham, G., Qiu, Y. & Inouye, M. FlashPCA2: principal component analysis of Biobank-scale genotype datasets. *Bioinformatics* **33**, 2776–2778 (2017).
44. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Methodol.* **57**, 289–300 (1995).
45. Van Greevenbroek, M. M. et al. The cross-sectional association between insulin resistance and circulating complement C3 is partly explained by plasma alanine aminotransferase, independent of central obesity and general inflammation (the CODAM study). *Eur. J. Clin. Invest.* **41**, 372–379 (2011).
46. Tigchelaar, E. F. et al. Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics. *BMJ Open* **5**, e006772 (2015).
47. Hofman, A. et al. The Rotterdam Study: 2014 objectives and design update. *Eur. J. Epidemiol.* **28**, 889–926 (2013).
48. Skyler, J. S. Pulmonary insulin update. *Diabetes Technol. Ther.* **7**, 834–839 (2005).
49. Storey, J. D. A direct approach to false discovery rates. *J. R. Stat. Soc. Series B Methodol.* **64**, 479–498 (2002).
50. Alexa, A. & Rahnenfuhrer, J. topGO: enrichment analysis for gene ontology. R package version 2 (2010).
51. Lippert, C., Casale, F. P., Rakitsch, B. & Stegle, O. LIMIX: genetic analysis of multiple traits. *bioRxiv* <https://doi.org/10.1101/003905> (2014).

In the format provided by the authors and unedited.

# A linear mixed-model approach to study multivariate gene–environment interactions

Rachel Moore<sup>1,2,3,9</sup>, Francesco Paolo Casale<sup>4,9</sup>, Marc Jan Bonder<sup>2</sup>, Danilo Horta<sup>2</sup>, BIOS Consortium<sup>5</sup>, Lude Franke<sup>6</sup>, Inês Barroso<sup>1\*</sup> and Oliver Stegle<sup>2,7,8\*</sup>

<sup>1</sup>Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. <sup>2</sup>European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK. <sup>3</sup>University of Cambridge, Cambridge, UK. <sup>4</sup>Microsoft Research New England, Cambridge, Massachusetts, USA. <sup>5</sup>A full list of members and affiliations appears at the end of the paper. <sup>6</sup>University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the Netherlands. <sup>7</sup>European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany.

<sup>8</sup>Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany.

<sup>9</sup>These authors contributed equally: Rachel Moore, Francesco Paolo Casale. \*e-mail: [ib1@sanger.ac.uk](mailto:ib1@sanger.ac.uk); [oliver.stegle@embl.de](mailto:oliver.stegle@embl.de)



## Supplementary tables

Main simulations										
$N$	-	-	-	1,000	2,000	<b>5,000</b>	-	-	-	-
$L$	-	2	10	20	40	<b>60</b>	80	100	-	-
$\pi$	-	-	3.3%	16.7%	33.3%	<b>50%</b>	66.7%	100%	-	-
$v_g$	-	-	-	-	-	<b>0.6%</b>	-	-	-	-
$\rho$	-	0.0	0.1	0.3	0.5	<b>0.7</b>	0.8	0.9	1.0	-
$v_e$	-	-	-	0	0.1	<b>0.2</b>	0.3	0.4	-	-
$v_{pop}$	-	-	-	-	-	<b>0.4</b>	-	-	-	-
$L_{unobs}$	-	-	-	-	-	<b>0</b>	10	20	30	40
Heritable environments										
$r^2$	0-0.1	0.1-0.2	0.2-0.5	0.5-0.8	0.8-1	<b>1</b>	-	-	-	-
$v_x$	-	0.01	0.02	0.05	0.10	<b>0.20</b>	-	-	-	-
Skewed (Gamma-distributed) environments										
$a$	-	100	5	3	2	1	0.5	0.2	0.1	-
Binary environments										
$f_B$	-	0	0.25	0.50	0.75	<b>1</b>	-	-	-	-
$\nu$	-	0.50	0.20	0.10	0.05	<b>0.02</b>	0.01	0.005	-	-

**Supplementary Table 1 | Parameters used for the simulation experiments.** Shown are the default parameter values and ranges as considered in the calibration and power experiments. Specifically, simulation experiments shown in (**Fig. 1,2, Supp. Fig. 2**) are based on empirical environments from UK Biobank, varying the sample size ( $N$ ), the number of environments ( $L$ ), the percentage of environments that contribute to GxE ( $\pi$ ), sample variance explained by the total genetic effect (G+GxE effect,  $v_g$ ), fraction of the genetic variance due to GxE ( $\rho$ ) and the variance explained by additive environmental effects ( $v_e$ ), population structure ( $v_{pop}$ ), residual noise ( $v_n$ ) and number of unobserved environments contributing to GxE ( $L_{unobs}$ ). Default parameter values are shown in bold and are left constant when varying other parameters. In additional simulations based on synthetic environments (**Supp. Fig. 3,4**), we vary LD between variants that affect environments and variants with G or GxE effects ( $r^2$ ), the average fraction of variance explained by the variant across environments ( $v_x$ , heritability). For simulated skewed environments, we vary the shape parameter of the gamma distribution ( $a$ ) and finally, for binary environments, we vary both the fraction of environments that are binary ( $f_B$ ) and the event frequency ( $\nu$ ). All remaining genetic parameters were set to default parameters.

Method name	GxE	Additive E	Test type	Number parameters for additive E	Number DoF Interaction test	Number DoF association test
Multi-environment tests						
<b>StructLMM</b>	<b>Random</b>	<b>Random</b>	<b>Score</b>	<b>1</b>	<b>1</b>	<b>2</b>
MultiEnv-Renv-LRT	Fixed	Random	LRT	1	60	61
MultiEnv-Fenv-LRT	Fixed	Fixed	LRT	60	60	61
MultiEnv-Renv-Score	Fixed	Random	Score	1	60	61
MultiEnv-Fenv-Score	Fixed	Fixed	Score	60	60	61
Single-environment tests						
<b>SingleEnv-Renv</b>	<b>Fixed</b>	<b>Random</b>	<b>LRT</b>	<b>1</b>	<b>1</b>	<b>2</b>
SingleEnv-Fenv	Fixed	Fixed	LRT	60	1	2
SingleEnv-Senv	Fixed	Fixed	LRT	1	1	2
Linear association tests						
<b>LMM-Renv</b>	<b>None</b>	<b>Random</b>	<b>LRT</b>	<b>1</b>	<b>NA</b>	<b>1</b>
LM-Fenv	None	Fixed	LRT	60	NA	1
LM	None	None	LRT	0	NA	1

**Supplementary Table 2 | Tabular overview of considered methods.** Shown is the name of the method (column 1; name of the association method displayed and ‘-int’ appended to the single- and multi-environment test names for the corresponding interaction tests), whether a random or fixed effect term is used to model GxE under the alternative hypothesis (column 2), whether a random or fixed effect term is used to model the additive environment (column 3), the statistical test used to assess the alternative hypothesis (column 4), how many parameters are used to model the additive environment (column 5), the number of additional parameters used to model the alternative hypothesis versus the null hypothesis for interaction and association tests (columns 6 and 7 respectively). The number of model parameters in the final three columns assume that 60 environmental variables are used in the test (default simulation setting; see **Supp. Table 1**). The tests are grouped into multi-environment tests, single-environment tests, and linear association tests. Representative methods that are considered in main text results are highlighted in bold.

**Supplementary Table 3 | Interactions identified by StructLMM for BMI in UK Biobank.** Provided as supplementary data file Supplementary\_table\_3.xlsx.

**Supplementary Table 4 | Associations identified by StructLMM and LMM in the association analysis of BMI using data from UK Biobank.** Provided as supplementary data file Supplementary\_table\_4.xlsx.

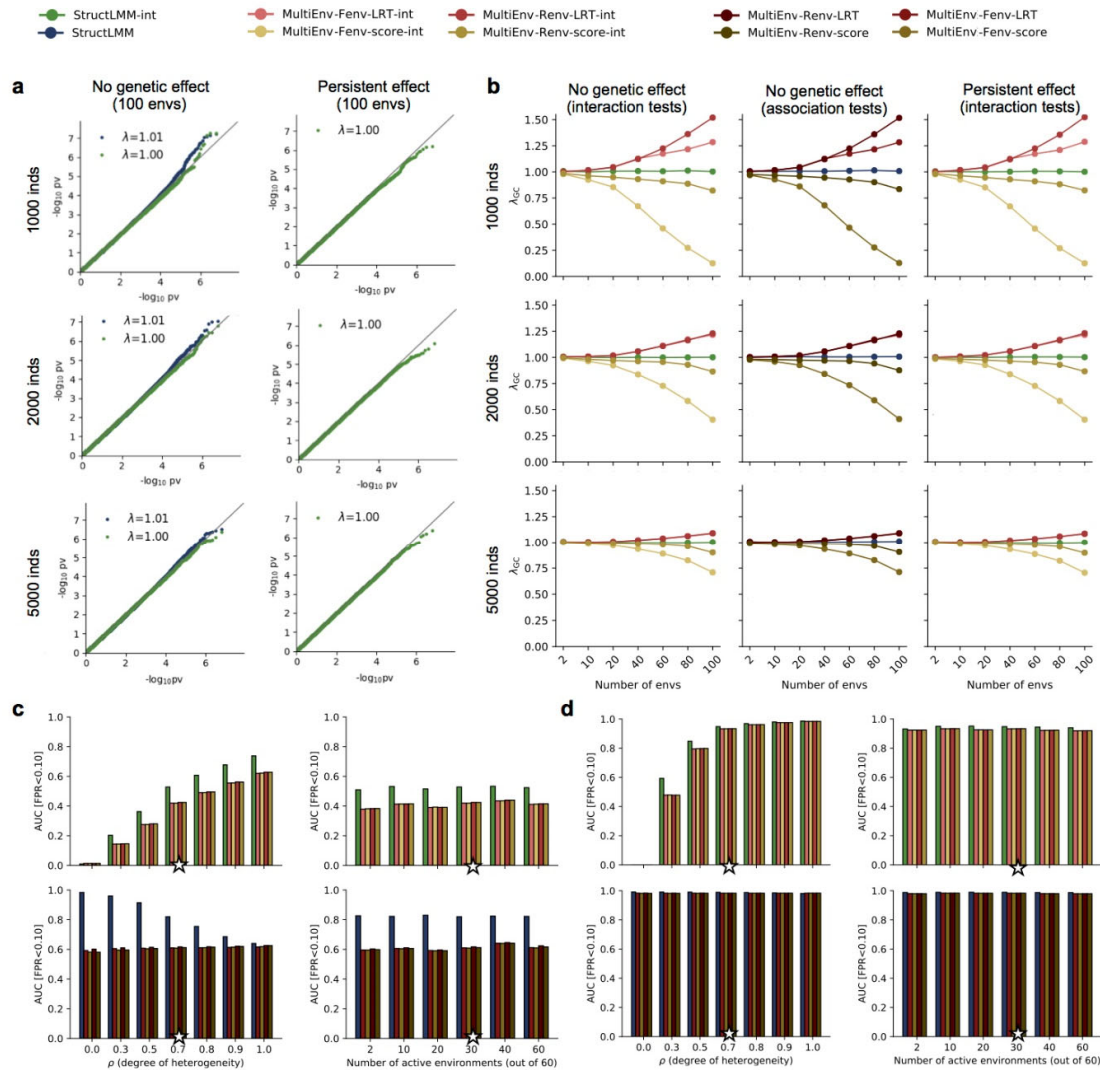
**Supplementary Table 5 | Summary table of interaction eQTL analysis in blood cohort.** Provided as supplementary data file Supplementary\_table\_5.xlsx.

**Supplementary Table 6 | Pathway enrichment analysis for interactions eQTL that are in linkage with GWAS loci.** Provided as supplementary data file Supplementary\_table\_6.xlsx.

**Supplementary Dataset 1 | eQTL Manhattan plots for interaction eQTL that colocalise with disease variants.** Manhattan plots provided as supplementary data file Supplementary\_data\_1.zip.

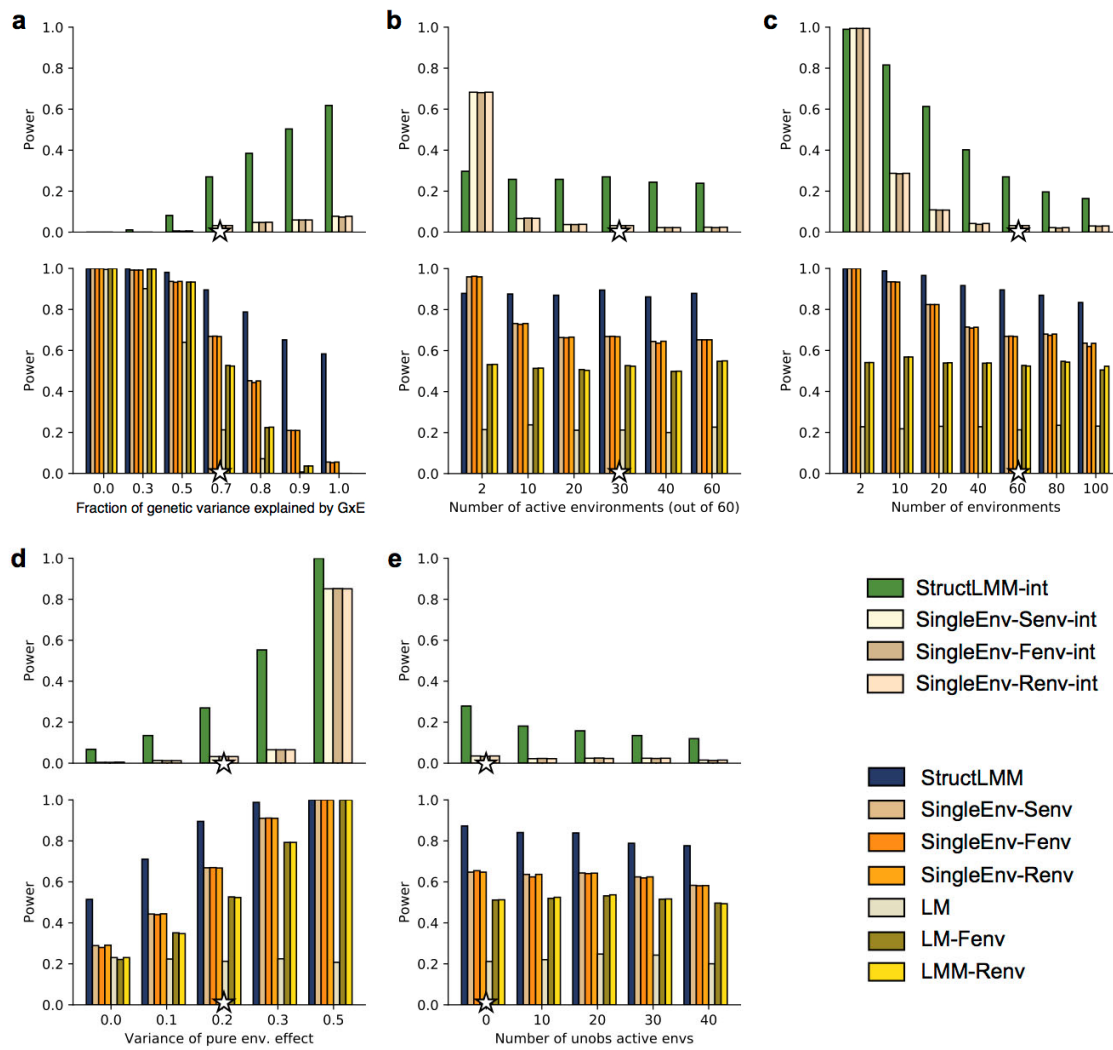
**Supplementary Dataset 2 | Interaction eQTL colocating with disease variants.** Figures analogous to **Fig. 5b-d** for 64 interaction eQTL with putative colocalisation with disease variants, provided as supplementary data file `Supplementary_data_2.zip`.

## Supplementary figures

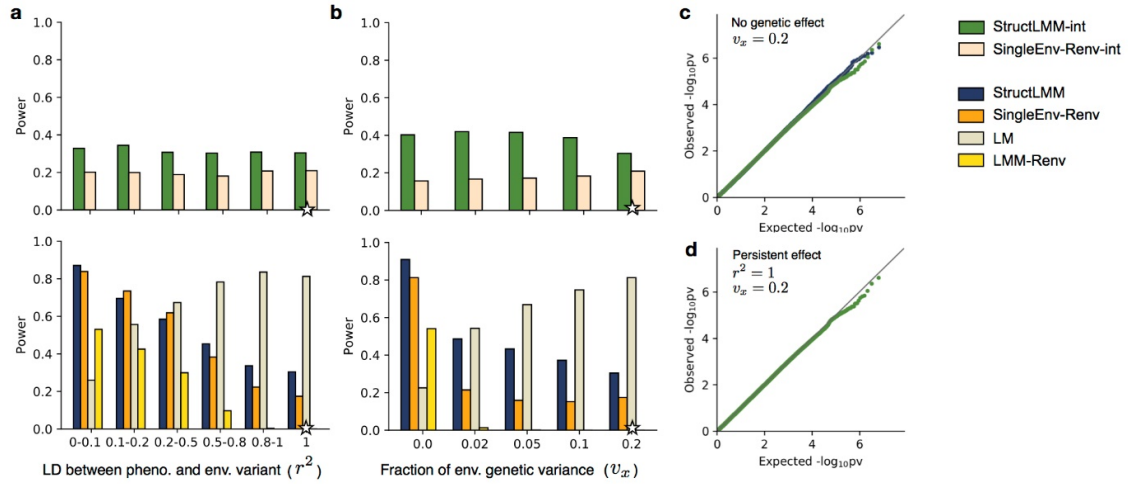


**Supplementary Figure 1 | Calibration of StructLMM and comparison with alternative multi-environment tests.** (a) QQ plots of negative log P values from StructLMM-int (green) and StructLMM (blue) for 103,527 variants on chromosome 21, either simulating no genetic effects (no G, no G $\times$ E,  $v_g = 0$ ) or simulating persistent genetic effects (no G $\times$ E,  $\rho = 0$ ; only StructLMM-int). Simulations are based on 100 environments ( $L = 100$ ,  $\pi = 100\%$ ). From top to bottom: increasing sample sizes of a synthetic population based on the European population from the 1000 Genomes project: 1,000 individuals, 2,000 individuals and 5,000 individuals. Default parameters were used for all other simulation settings; see **Supp. Table 1**.

(b) Genomic inflation factor  $\lambda_{GC}$  of P values from StructLMM and alternative multi-environment tests based on fixed effects (**Supp. Table 2**), for different numbers of environmental variables ( $L$ ,  $\pi = 100\%$ ; x-axis) and for increasing sample size ( $N$ , top to bottom). Shown are results from StructLMM-int and multi-environment fixed effect interaction tests (column 1), and equivalent association tests (column 2) when no genetic effects are simulated ( $v_g = 0$ ). Column 3 depicts results from StructLMM-int and multi-environment fixed effect interaction tests for simulated persistent genetic effects ( $\rho = 0$ ). Multi-environment fixed effect models using LR tests yielded inflated test statistics (inflation factors  $\lambda_{GC} > 1$ ) for large numbers of environmental factors in relation to the sample size, whilst score tests yielded deflated statistics (inflation factors  $\lambda_{GC} < 1$ ) for the corresponding settings. StructLMM was calibrated in all settings. (c,d) Performance assessment of alternative methods for detecting interactions (top panels) and associations (bottom panels) based on simulated data, using the settings as in **Fig. 2b,c** and **Supp. Fig. 2a,b** for two sample sizes:  $N = 2,000$  (c) and  $N = 5,000$  (d). Compared were StructLMM-int and alternative multi-environment interaction tests based on fixed effects (**Supp. Table 2**, top panels), StructLMM and additional multi-environment association tests (**Supp. Table 2**, bottom panels). As the fixed effect tests are not always calibrated (see panel b), shown are model performance values as assessed by the area under the curve (relative AUC, in the range  $0 < \text{FPR} < 0.10$ , normalised such that 0 corresponds to chance performance and 1 to an ideal model).

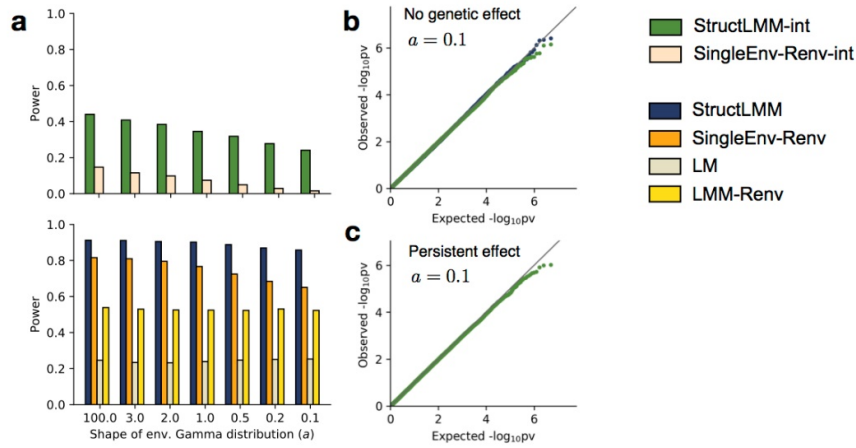


**Supplementary Figure 2 | Assessment of power using simulated data.** Power comparison of alternative methods for detecting interactions (top panels) and associations (bottom panels) based on simulated data, extending the results as shown in **Fig. 2**, varying (a) the fraction of the genetic variance explained by GxE ( $\rho$ ), (b) the number of environments with non-zero GxE effects ( $\pi$ ), (c) the total number of environments ( $L$ , 50% contributing to GxE effects), (d) the fraction of variance explained by additive environment effects ( $v_e$ ) and (e) the number of environments that contribute to GxE but are not used (observed,  $L_{unobs}$ ) for the respective tests. Considered were top panels: the StructLMM interaction test (StructLMM-int) and alternative implementations of single-environment interaction tests (**Supp. Table 2**); bottom panels: StructLMM association test and alternative implementations of 2-df fixed effect tests that jointly tests for persistent associations and interactions with a single environment (**Supp. Table 2**), as well as alternative implementations of linear (mixed) models to test for persistent effects (**Supp. Table 2**). Methods were assessed in terms of power (at FWER<1%) for detecting simulated causal variants (**Methods**). Stars denote default values of genetic parameters, which were retained when varying other parameters (see **Supp. Table 1 & Methods** for details on the simulation strategy). A synthetic European population of 5,000 individuals based on 1000 Genomes Project genotypes was used for all experiments (**Methods**).

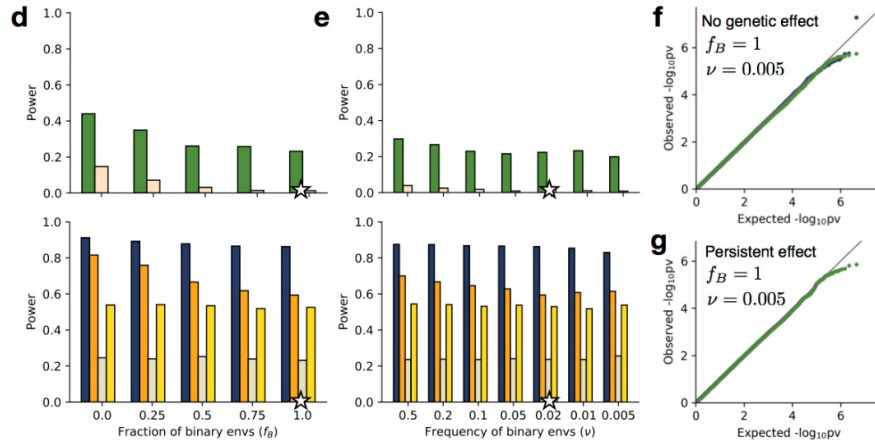


**Supplementary Figure 3 | Assessment of power and calibration in the presence of heritable exposures.** Power comparison of alternative methods for detecting interactions (top panels) and associations (bottom panels) with simulated genetic effects both on environmental variables and phenotypes. Considered were (a) increasing the LD ( $r^2$ ) between causal variants associated with the environment and those with additive G and GxE effects on the phenotype and (b) the average fraction of the exposures variance explained by genetic effects. Stars denote the default value that was retained when varying the other parameter (see **Supp. Table 1**). Top panel: power to detect interactions, considering the StructLMM interaction test (StructLMM-int) and the default implementation of the single-environment interaction tests (SingleEnv-Renv-int, **Supp. Table 2**). Bottom panel: power to detect association, considering StructLMM, the default 2-df fixed effect test that jointly tests for persistent associations and interactions with a single environment (SingleEnv-Renv, **Supp. Table 2**), and linear models to test for persistent effects (LM, LMM-Renv, **Supp. Table 2**). Models were assessed in terms of power (FWER<1%) for detecting simulated causal variants (**Methods**). (c) QQ plot of negative log P values obtained from StructLMM-int (green) and StructLMM (blue) for pronounced gene-exposure correlations ( $v_x = 0.2$ ), when no genetic effects are simulated ( $v_g = 0$ ) for 103,527 variants on chromosome 21. (d) QQ plot of negative log P values obtained from StructLMM-int (green) when persistent genetic effects were simulated ( $\rho = 0$ ), assuming strong LD between gene-exposure effects and variants associated with phenotype ( $v_x = 0.2, r^2 = 1$ ) for 103,527 variants on chromosome 21. Stars denote default values of genetic parameters, which were retained when varying other parameters (see **Supp. Table 1 & Methods** for details on the simulation strategy). A synthetic European population of 5,000 individuals based on 1000 Genomes Project genotypes was used for all experiments (**Methods**).

### Skewed (Gamma-distributed) environments

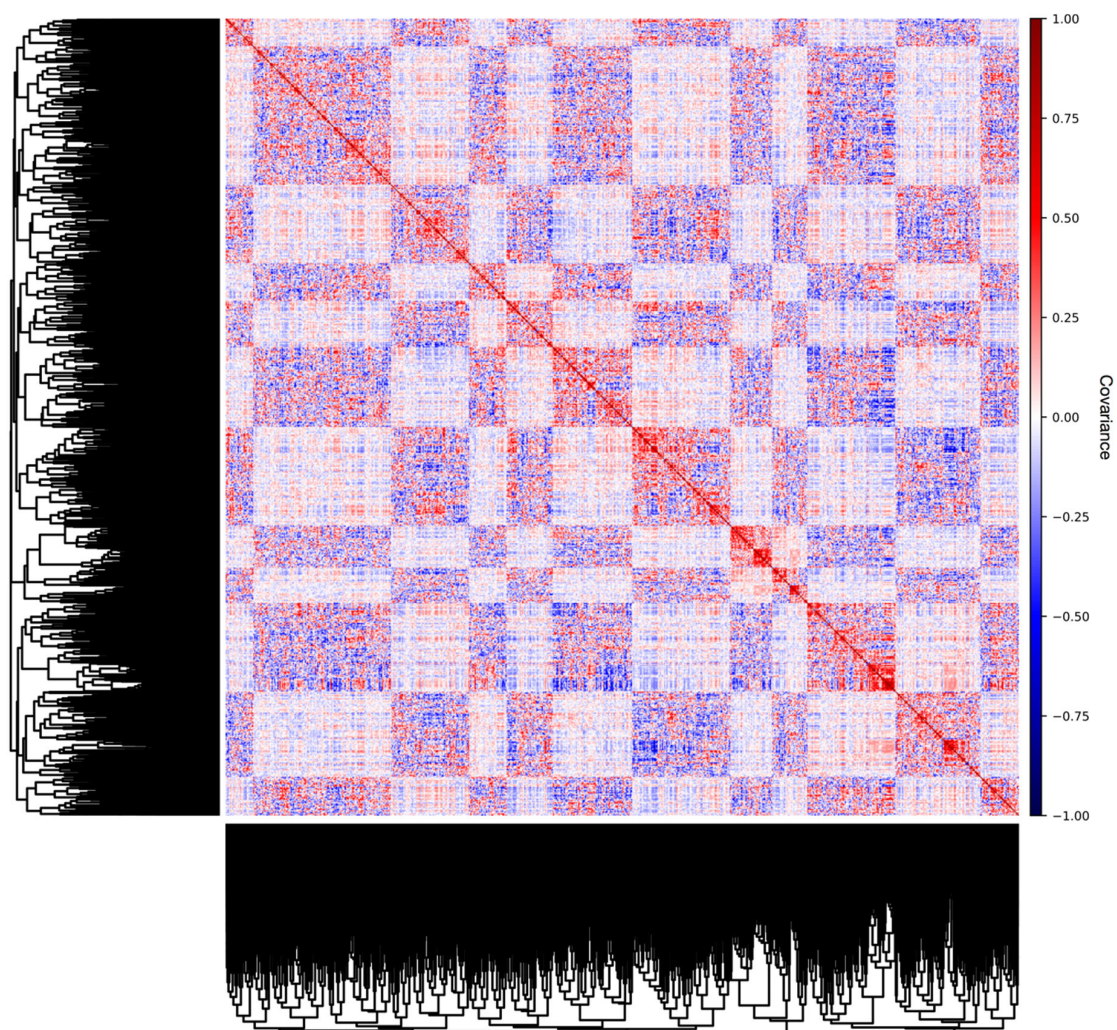


### Binary Environments



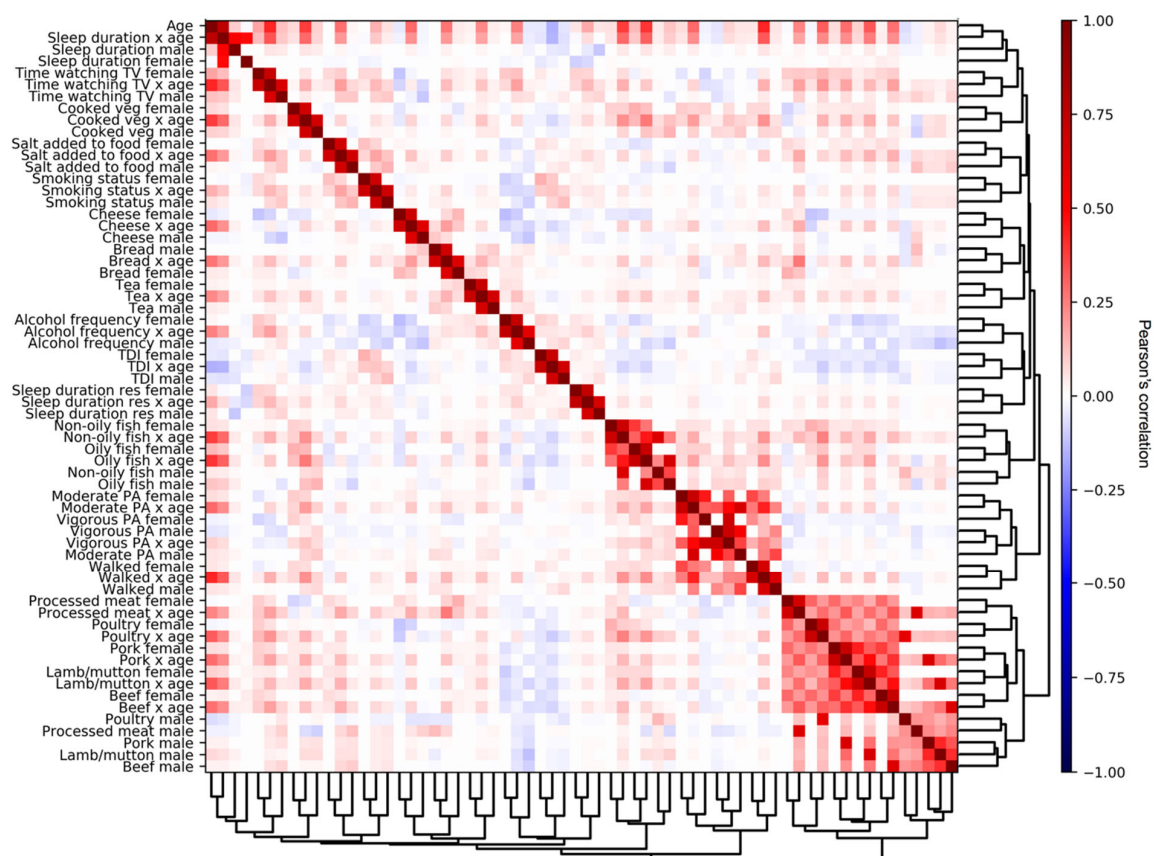
**Supplementary Figure 4 | Assessment of power and calibration using simulated data for skewed and binary environments.** (a-c) Power comparison and calibration when simulating skewed environments drawn from a Gamma distribution. (a) Power comparison for varying shape of the gamma distribution (low values correspond to skewed environments, shape 100 correspond to approximately Gaussian distributed environments). (b,c) QQ plots of negative log P values obtained from StructLMM and StructLMM-int on data with skewed environments for 103,527 variants on chromosome 21, when simulating no genetic effect ( $v_g = 0$ ) (b) or when simulating persistent genetic effects (c, StructLMM-int only,  $\rho = 0$ ). (d-g) Power comparison and calibration for simulated binary environments. (d) Power comparison when varying the fraction of binary environments  $f_B$  (for constant event frequency  $\nu$ ). (e) Power comparison when varying the event frequency of binary environments. Stars denote the default value that was retained when varying the other parameter (see **Supp. Table 1**). (f,g) QQ plots of negative log P values obtained from StructLMM and StructLMM-int for rare binary environments for 103,527 variants on chromosome 21, either simulating no genetic effect ( $v_g = 0$ ) (f) or for simulated persistent genetic effects (g, StructLMM-int only,  $\rho = 0$ ). Fixed parameter settings are indicated in the top left corner of the corresponding panels. A synthetic European population of 5,000 individuals based on 1000 Genomes Project genotypes was used for all experiments (**Methods**).



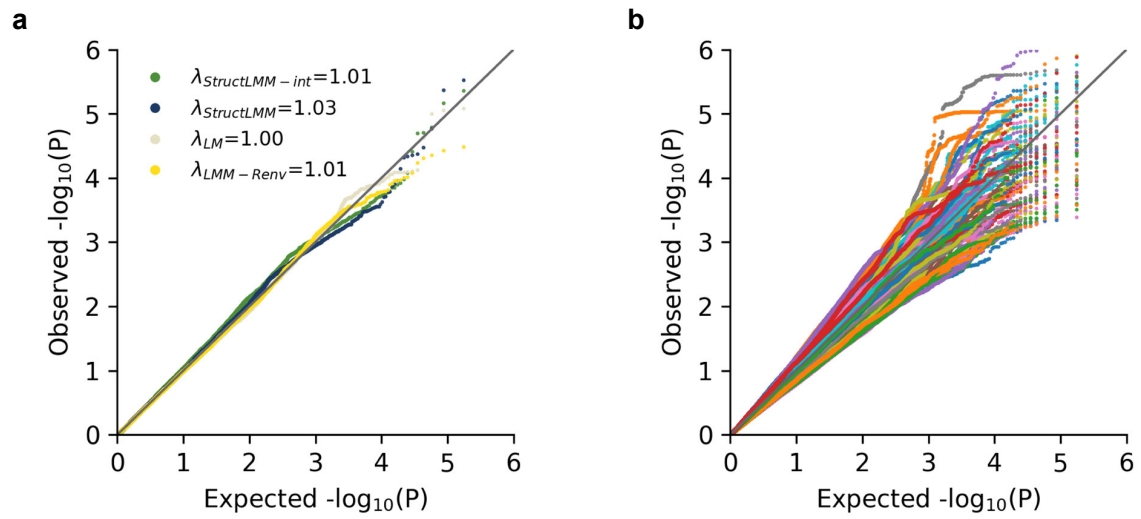


**Supplementary Figure 5 | Covariance structure of UK Biobank individuals based on 64 environmental variables.** Sample covariance matrix for 5,000 randomly selected individuals, calculated based on 64 environmental variables considered for UK Biobank analyses: 12 diet-related factors, three factors linked to physical activity and a six lifestyle factors, modelled as gender-adjusted and age-adjusted (**Methods**). Dark red denotes pairs of individuals with stronger environmental similarity, whilst blue corresponds to negative covariance of environmental similarity (anti correlation). Blocks of strong correlation/anti correlation, correspond to groups of individuals of the same gender.

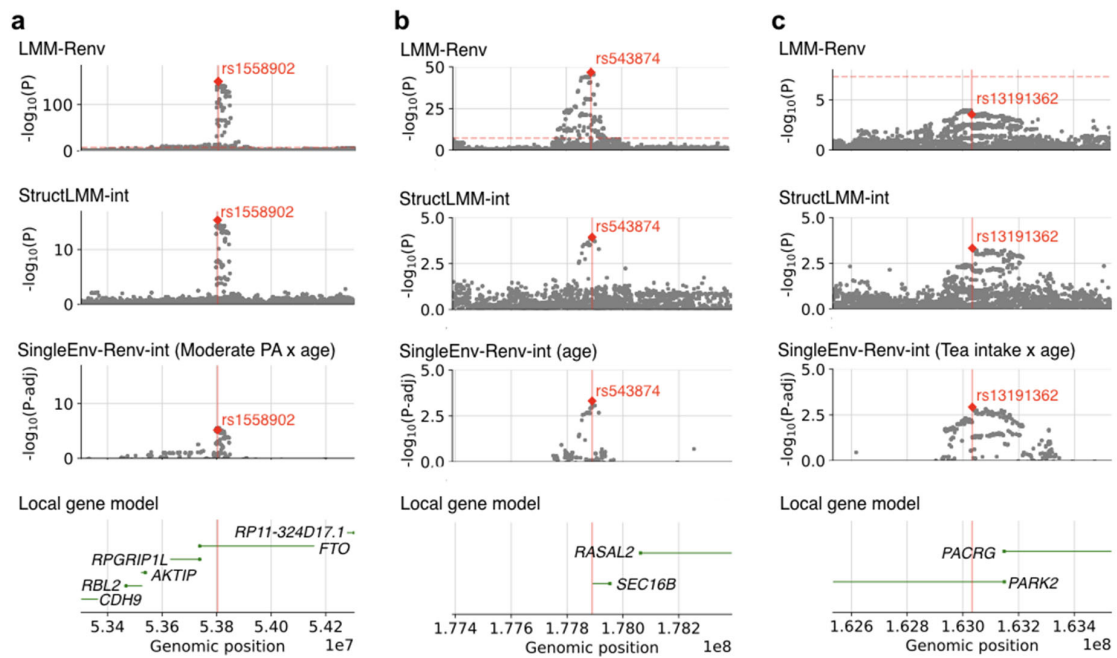




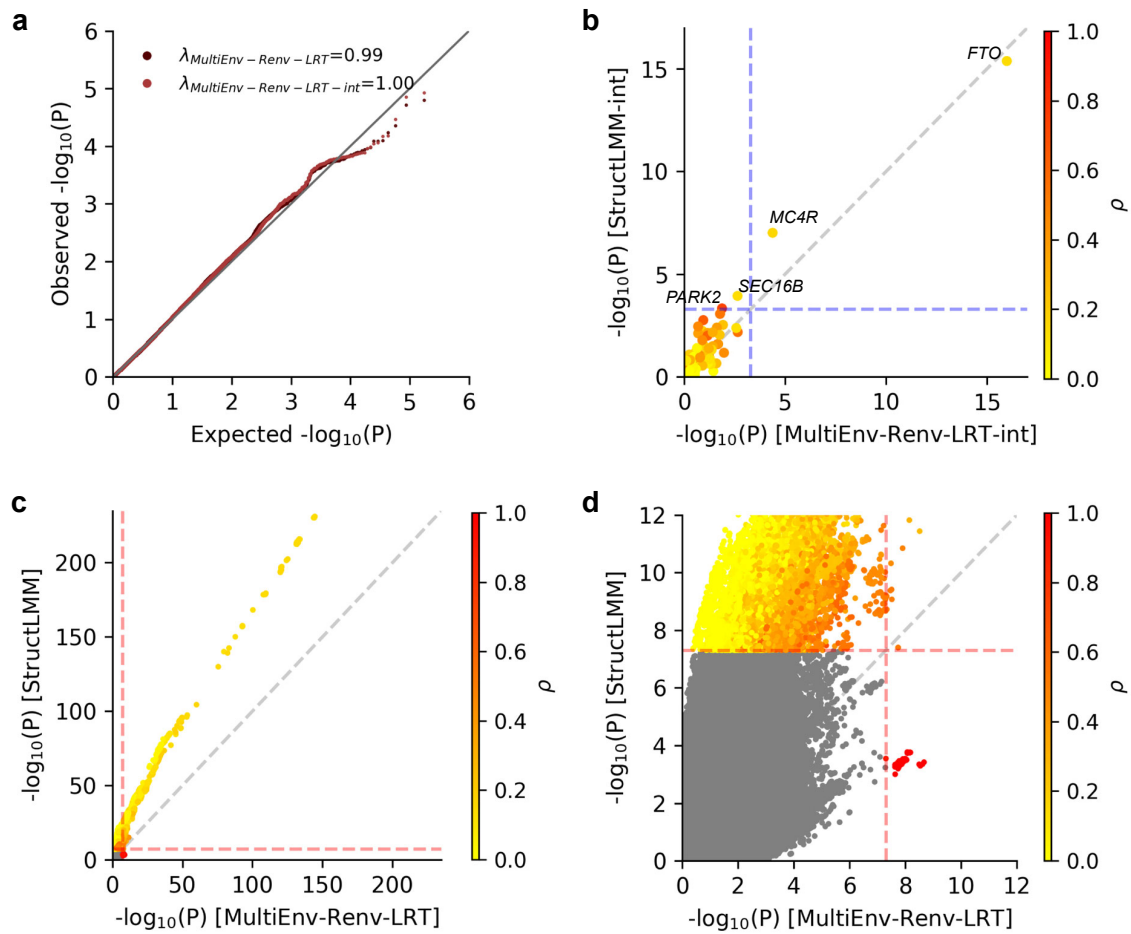
**Supplementary Figure 6 | Structure of 64 environmental variables considered for UK Biobank analyses.** Shown are correlation coefficients between pairs of environmental variables, considering 12 diet-related factors, three factors linked to physical activity and a six lifestyle factors, modelled as gender-adjusted and age-adjusted (**Methods**). Environments are ordered using hierarchical clustering.



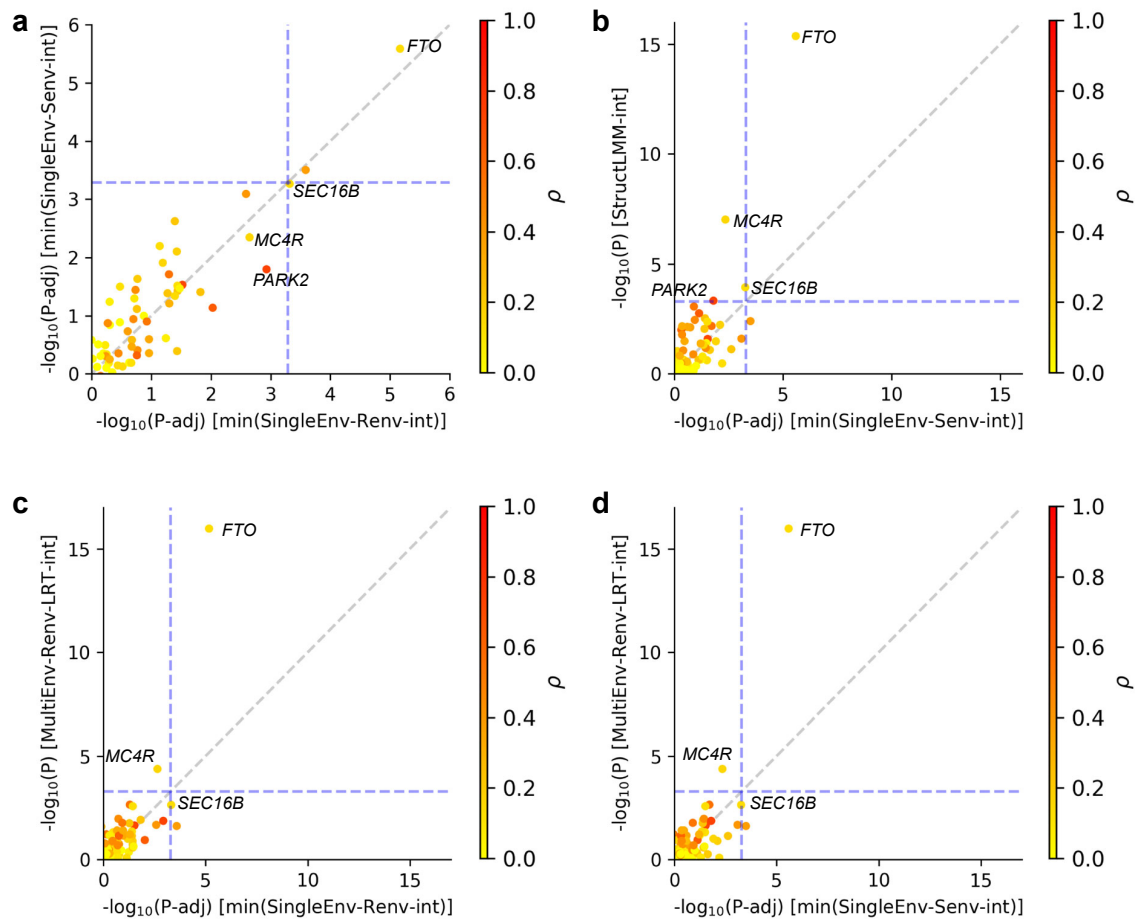
**Supplementary Figure 7 | Calibration of interaction and association tests on UK Biobank data.** QQ plots of negative log P values from different interaction and association tests applied to UK Biobank BMI phenotype data ( $n = 252,188$  unrelated individuals of European ancestry) based on permuted genetic variants (chromosome 20, 173,297 variants). **(a)** QQ plots of negative log P values for StructLMM-int (green), StructLMM (blue), LM (light grey) and LMM-Renv (yellow). **(b)** QQ plots of negative log P values for SingleEnv-Renv-int, for each of 64 considered environmental variables. LM, LMM-Renv and StructLMM tests were calibrated, whereas the fixed effect interaction tests show variable levels of statistical calibration. See **Supp. Table 2** for an overview of considered methods.



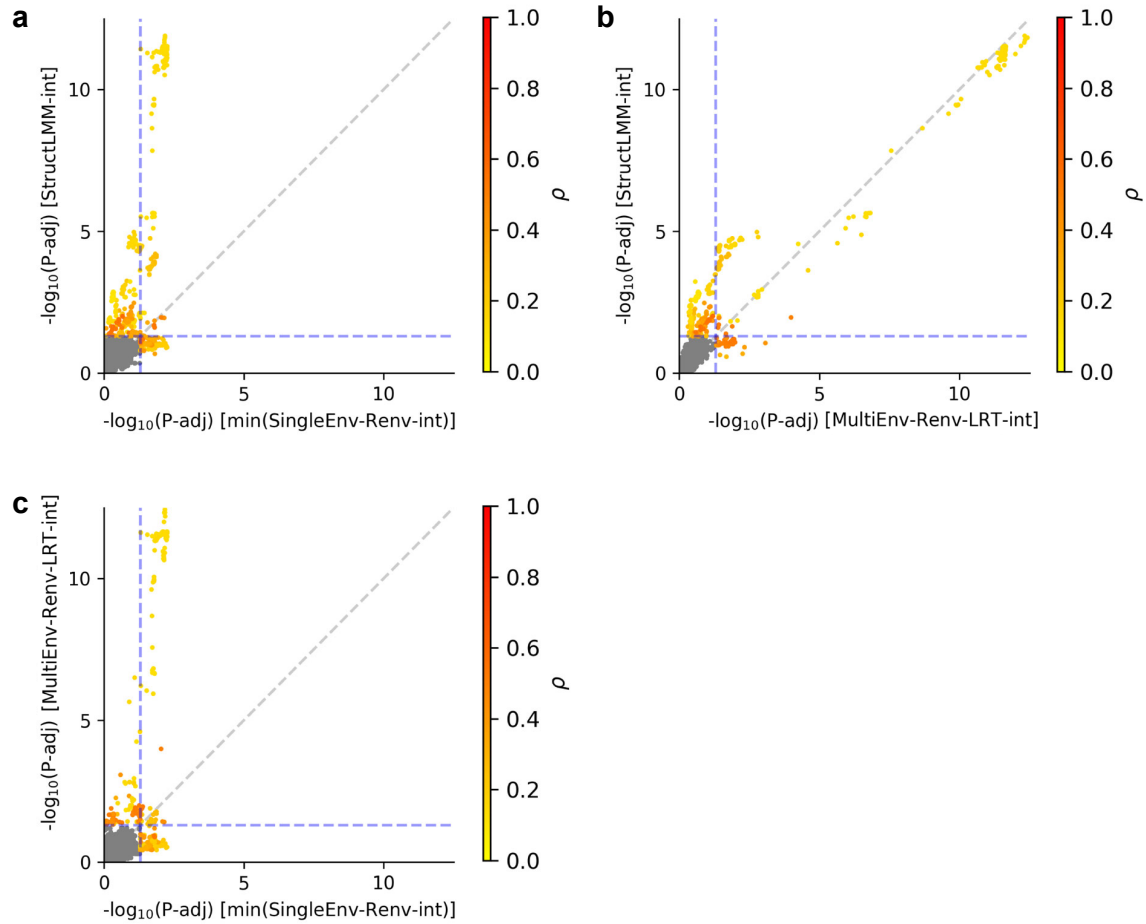
**Supplementary Figure 8 | Supplementary results for interactions identified by StructLMM.** (a-c) Local Manhattan plots of interactions identified by StructLMM at *FTO*, *SEC16B* and *PARK2* respectively. From top to bottom: LMM-Renv association test, StructLMM interaction test, SingleEnv-Renv interaction test for the environment with the most significant GxE effect at the respective GIANT variant, local gene models. The red vertical line indicates the position of the GIANT variant as in **Fig. 3a**.



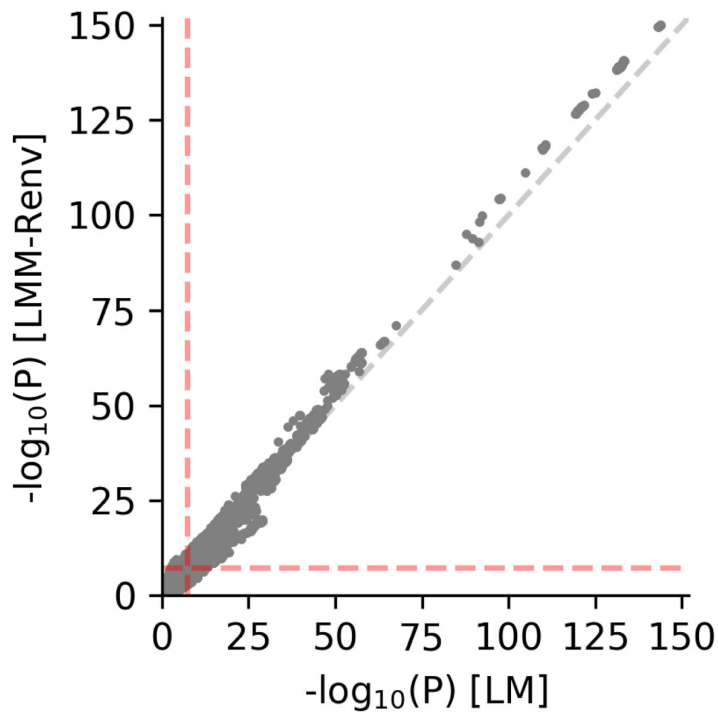
**Supplementary Figure 9 | Comparison of multi-environment GxE tests based on fixed effects on UK biobank data.** (a) QQ plots of negative log P values from a multi-environment fixed effect model to test for associations while accounting for heterogeneity in effect sizes due to GxE (MultiEnv-Renv-LRT) and an interaction test (MultiEnv-Renv-LRT-int) applied to UK Biobank BMI phenotype data and permuted genetic variants (chromosome 20, 173,297 variants), which are calibrated for this sample size ( $n = 252,188$ ). (b) Scatter plot of negative log P values from GxE interaction tests at 97 GIANT variants (Locke *et al.*, 2015), considering the MultiEnv-Renv-LRT-int interaction test (x-axis) versus the StructLMM-int test (y-axis). Dashed lines correspond to  $\alpha < 0.05$ , Bonferroni adjusted for the number of tested GIANT variants and colour denotes the estimated fraction of genetic variance due to GxE (fitted parameter  $\rho$ ). StructLMM-int and MultiEnv-Renv-LRT-int identified four and two significant loci respectively. (c) Scatter plot of negative log P values from the MultiEnv-Renv-LRT association test (x-axis) versus the StructLMM association test (y-axis). Dashed lines indicate genome-wide significance at  $P < 5 \times 10^{-8}$  and colour denotes the estimated fraction of genetic variance due to GxE (fitted parameter  $\rho$ ), with (d) displaying a zoom-in view of variants close to genome-wide significance. StructLMM and MultiEnv-Renv-LRT identify 17,630 and 2,037 significant variants respectively. All displayed results are generated using 252,188 unrelated individuals of European ancestry.



**Supplementary Figure 10 | Comparison of different single and multi-environment interaction test based on fixed effects on UK Biobank data.** Compared are two alternative single-environment tests (SingleEnv-Renv-int, SingleEnv-Senv-int), as well as a multi-environment test (MultiEnv-Renv-LRT-int) based on fixed effects and StructLMM-int at 97 GIANT variants (Locke *et al.*, 2015). Scatter plot of negative log P values of (a) fixed effect tests, either accounting for additive environment effects of all environments (SingleEnv-Renv-int, using random effects, x-axis) or of the single environment that is tested (SingleEnv-Senv-int, using fixed effects, y-axis), (b) fixed effect tests, that account for additive environmental effects of the single environment tested (SingleEnv-Senv-int, x-axis) versus StructLMM-int (y-axis), (c) single-environment fixed effect interaction test with an additive environmental random effect component (SingleEnv-Renv-int, x-axis) versus the multi-environment fixed effect interaction test (MultiEnv-Renv-LRT-int, y-axis) and (d) single-environment fixed effect interaction test with a fixed effect additive environment term accounting for the single environment that is tested for GxE effects under the alternative (SingleEnv-Senv-int, x-axis) versus the multi-environment fixed effect interaction test (MultiEnv-Renv-LRT-int, y-axis). Dashed lines correspond to  $\alpha < 0.05$  and colour denotes the estimated fraction of genetic variance due to GxE (fitted parameter  $\rho$ ). For single-environment models (SingleEnv-Renv-int, SingleEnv-Senv-int), shown is the Bonferroni adjusted minimum P value across the set of tested environments (P-adj). SingleEnv-Renv-int and SingleEnv-Senv-int identify the same significant loci, but different significant loci to those identified by MultiEnv-Renv-LRT-int. All displayed results are generated using 252,188 unrelated individuals of European ancestry.

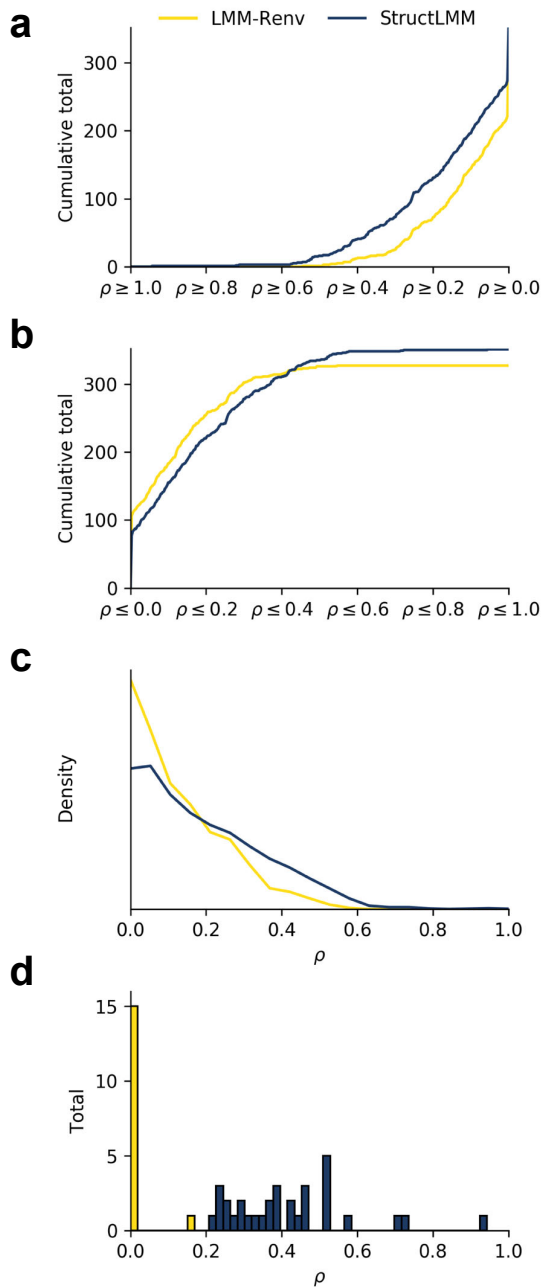


**Supplementary Figure 11 | Comparison of interaction test results for genome-wide variants associated with BMI on UK Biobank data.** Comparison of the StructLMM interaction test (StructLMM-int), a single-environment interaction test (SingleEnv-Renv-int) as well as multi-environment fixed effect interaction test (MultiEnv-Renv-LRT-int), at 17,606 variants that were significantly associated with BMI based on an LMM (LMM-Renv,  $P < 5 \times 10^{-8}$ ). Scatter plots of negative log Benjamini-Hochberg adjusted P values (P-adj) between (a) SingleEnv-Renv-int (x-axis) versus StructLMM-int (y-axis), (b) MultiEnv-Renv-LRT-int (x-axis) versus StructLMM-int (y-axis) and (c) SingleEnv-Renv-int (x-axis) versus MultiEnv-Renv-LRT-int (y-axis). Dashed lines correspond to  $P\text{-adj} < 0.05$  and colour denotes the estimated fraction of genetic variance due to GxE (fitted parameter  $\rho$  based on StructLMM). For the single-environment test (SingleEnv-Renv-int), shown is the Bonferroni adjusted minimum P value across the tested environments (P-adj). StructLMM-int identified 23 loci with GxE, followed by SingleEnv-Renv-int (11 loci) and MultiEnv-Renv-LRT-int (9 loci; FDR < 5%, Benjamini-Hochberg adjusted, LD clumped loci,  $r^2 < 0.1$  within 500kb, **Methods**; see **Supp. Table 3** for full results). All displayed results are generated using 252,188 unrelated individuals of European ancestry.



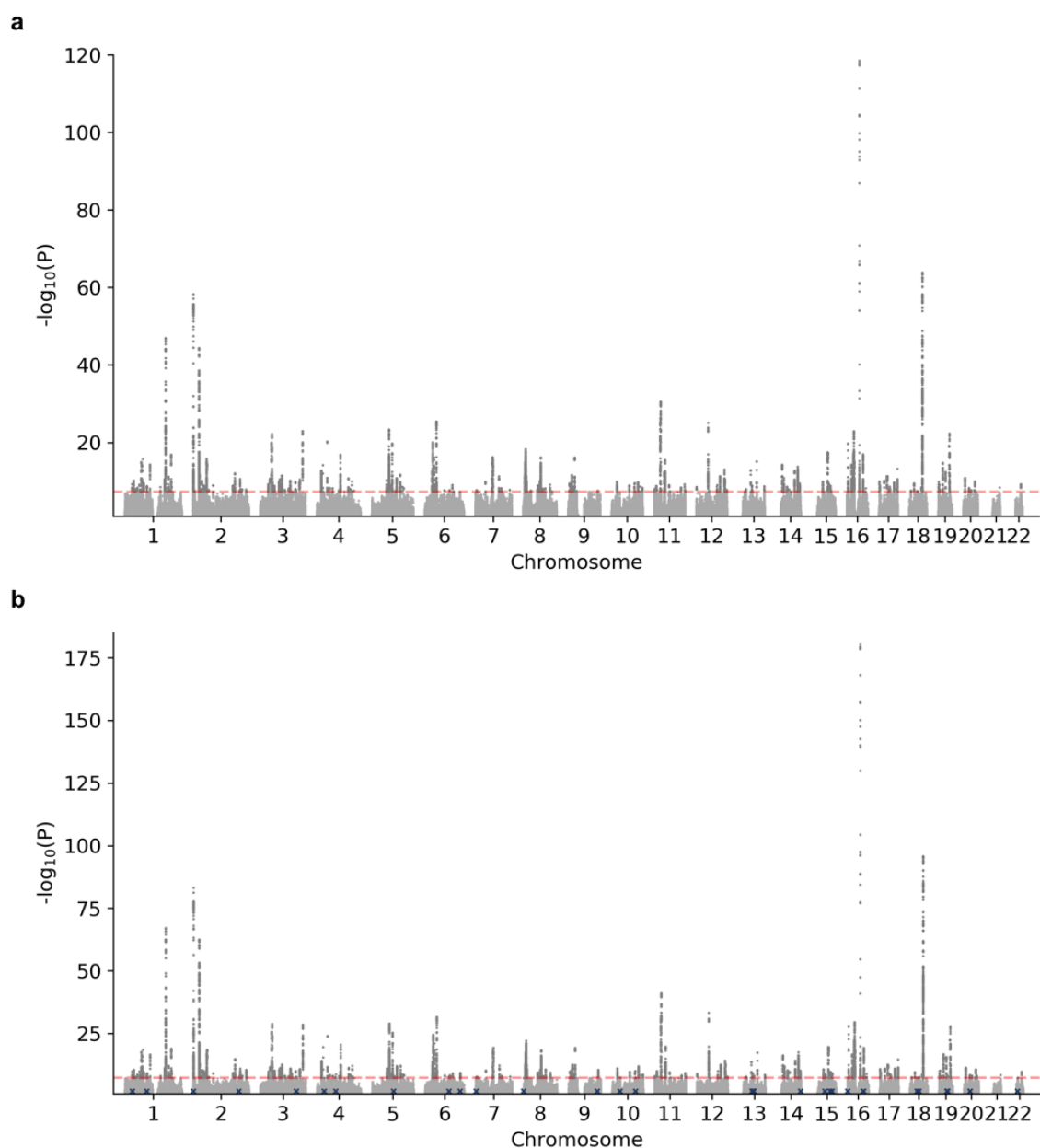
**Supplementary Figure 12 | Comparison of LMM and LM results on UK Biobank data.**

Scatter plot of genome-wide negative log P values from the LM association test, without accounting for additive environmental effects (x-axis), versus an LMM association test that accounts for additive environmental effects using the same random effect component as used in StructLMM (LMM-Renv, y-axis). Dashed lines indicate genome-wide significance at  $P < 5 \times 10^{-8}$ . All displayed results are generated using 252,188 unrelated individuals of European ancestry.

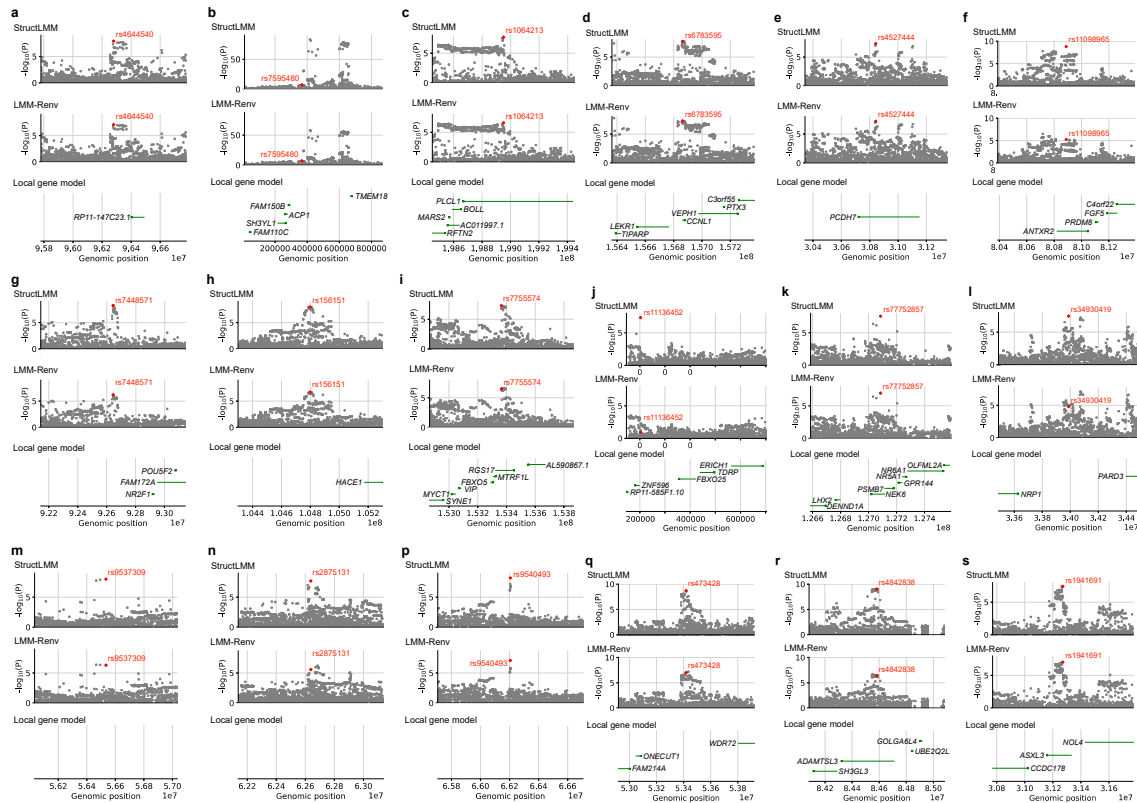


**Supplementary Figure 13 | Distribution of the estimated extent of GxE for significant loci identified by the StructLMM association test and LMM-Renv on UK Biobank data.** Cumulative number of significant associations ( $P < 5 \times 10^{-8}$ , LD clumped loci,  $r^2 < 0.1$  within 500kb, **Methods**) identified by LMM-Renv (yellow,  $N = 327$  loci) and StructLMM (blue,  $N = 351$  loci) in decreasing (a) and increasing (b) order of the estimated extent of GxE (fitted parameter  $\rho$ ). (c) Estimated density of the fraction of genetic variance due to GxE ( $\rho$ ) for loci identified by LMM-Renv (yellow) and StructLMM (blue). (d) Histogram of the fraction of genetic variance due to GxE ( $\rho$ ), considering the subset of loci exclusively identified by either approach (**Methods**): LMM-Renv (yellow, total 16), StructLMM (blue, total 32). Loci identified by StructLMM tended to have a greater extent of GxE, whereas loci that were LMM-Renv specific tended to have no or little evidence for effect size heterogeneity due to GxE. The latter can be explained by the fact that StructLMM will be slightly less powered than the LMM-Renv when no or very little GxE is present since StructLMM is penalised for testing multiple values  $\rho$  (see also **Supp. Fig. 2**).

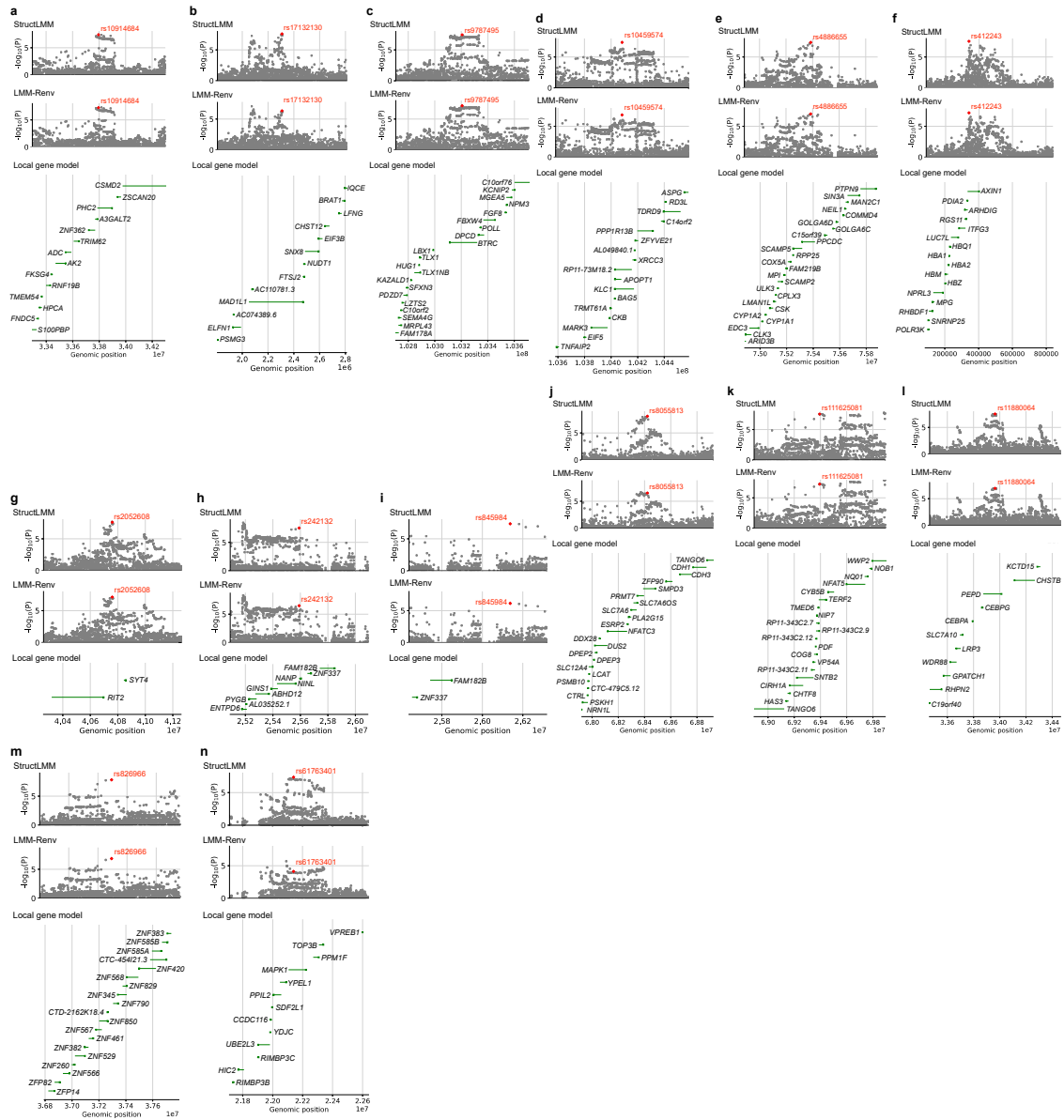




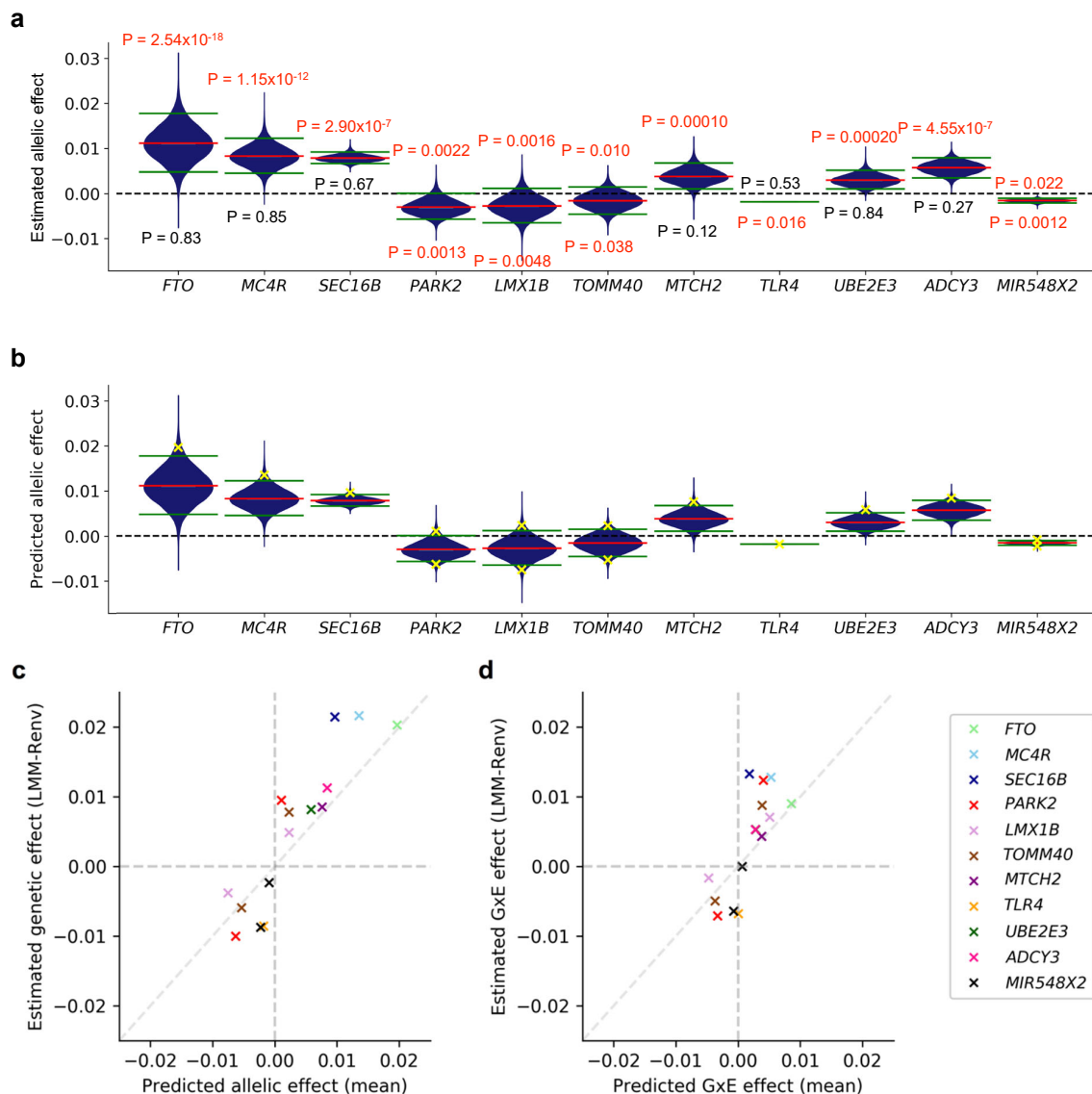
**Supplementary Figure 14 | Results from association tests applied to UK Biobank data.** Manhattan plots, showing negative log P values obtained from **(a)** LMM-Renv and **(b)** the StructLMM association test. Additional loci identified by StructLMM ( $P < 5 \times 10^{-8}$ ) as in **Supp. Table 4** are highlighted with a blue cross. The dashed red line denotes the genome-wide significance threshold ( $P < 5 \times 10^{-8}$ ). All displayed results are generated using 252,188 unrelated individuals of European ancestry.



**Supplementary Figure 15 | Local Manhattan plots for additional association loci identified by StructLMM versus LMM-Renv on UK Biobank data.** Shown are additional associations as in **Supp. Table 4**, with the red vertical line indicating the position of the StructLMM lead variant. From top to bottom: Manhattan plot of negative log P values from StructLMM, LMM-Renv and the local gene models for **(a-s)** *RP11-147C23.1*, *FAM150B*, *PLCL1*, *CCNL1*, *PCDH7*, *ANTXR2*, *NR2F1*, *HACE1*, *RGS17*, *ZNF596*, *NEK6*, *NRP1*, rs9537309 (no gene within +/-500 kb), rs2875131 (no gene within +/-500 kb), rs9540493 (no gene within +/-500 kb), *ONECUT1*, *ADAMTSL3* and *ASXL3* respectively. Associations were assigned to genes based on the nearest protein coding gene within +/-500 kb respectively. The maximum LD ( $r^2$ ) between significant StructLMM variants in the loci exclusive to StructLMM under consideration and significant LMM-Renv variants in other loci within the +/-500 kb region is 0.052, 0.00049 for panels **(b)** and **(d)** respectively. All displayed results are generated using 252,188 unrelated individuals of European ancestry.



**Supplementary Figure 16 | Local Manhattan plots for additional association loci identified by StructLMM versus LMM-Renv on UK Biobank data.** Shown are additional associations as in **Supp. Table 4**, with the red vertical line indicating the position of the StructLMM lead variant. From top to bottom: Manhattan plot of negative log P values from StructLMM, LMM-Renv and the local gene models. **(a-n)** *PHC2*, *MAD1L1*, *BTRC*, *RP11-73M18.2*, *PPCDC*, *AXIN1*, *RIT2*, *NANP*, *FAM182B*, *SMPD3*, *TERF2*, *PEPD*, *ZNF790* and *MAPK1* respectively. Loci were assigned to genes based on the nearest protein coding gene within +/-500 kb respectively. The maximum LD ( $r^2$ ) between significant StructLMM variants in the loci exclusive to StructLMM under consideration and significant LMM variants in other loci within the +/-500 kb region is 0.054 for panel **(h)**. All displayed results are generated using 252,188 unrelated individuals of European ancestry.

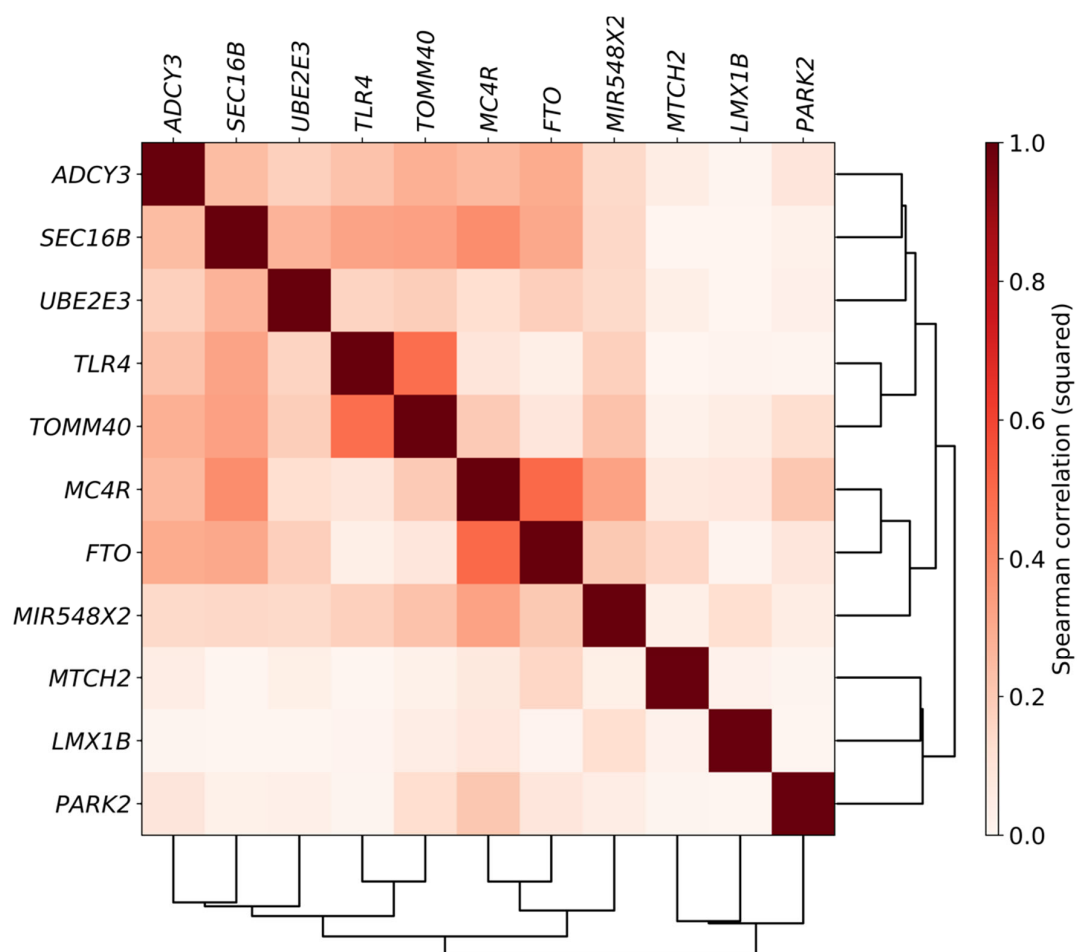


### Supplementary Figure 17 | Out of sample prediction of per-individual allelic effect sizes.

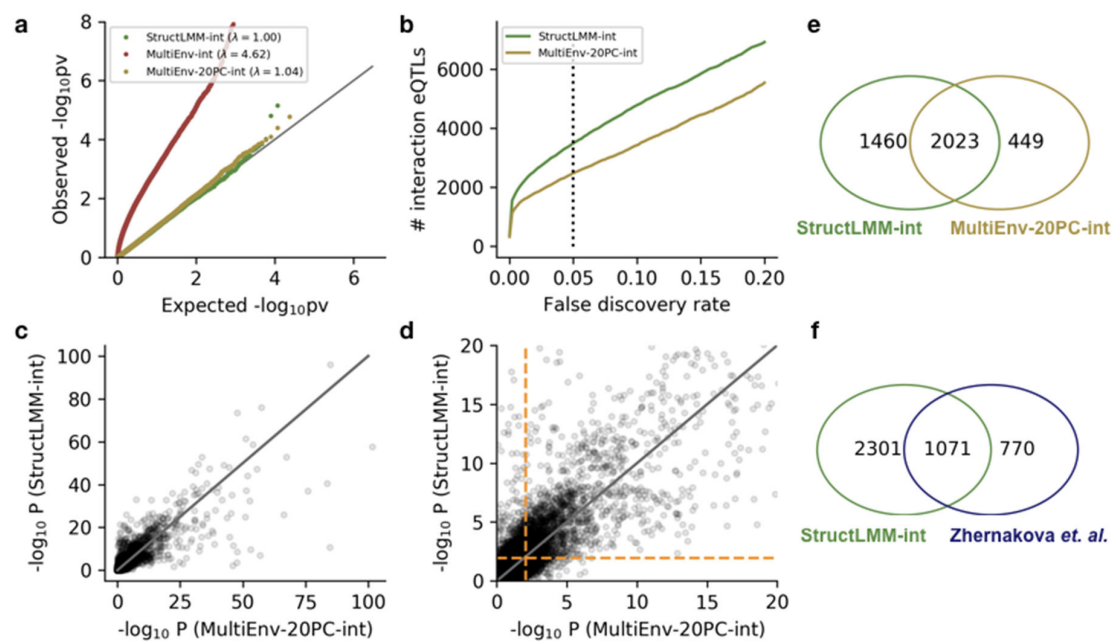
Assessment of per-individual allelic effects for 11 GIANT variants with evidence for GxE (lenient Benjamini-Hochberg adjustment,  $FDR < 0.05$ , **Supp. Table. 3**), considering a 50:50 split of the cohort into training and test fractions. **(a)** Violin plots, displaying the estimated density of individuals in the training fraction ( $n = 126,094$ ) that exert an allelic effect of a particular size given the distribution of the environmental factors within the population (in sample estimates as in **Fig. 4a**; see **Methods**). Mean and the top and bottom 5% strata of the effect size distribution are indicated by the red and green bars, respectively. P values denote the significance of genetic effects assessed using LMM-Renv (not accounting for GxE) using individuals within the respective strata. Nominally significant associations ( $P < 0.05$ , two-sided LR test) highlighted in red. **(b)** Analogous allelic effect size distribution as in **a**, displaying predictions in the test fraction (out-of-sample predictions, based on environmental states of individuals in the test fraction ( $n = 126,094$ ) but without using BMI phenotypes; **Methods**). Yellow crosses denote the mean predicted genetic effect within the top and bottom 5% strata (considering strata with nominally significant associations in training fractions, i.e. P values that are highlighted in red in **a**). **(c)** Scatter plot of allelic effect sizes, predicted out of sample

for test set individuals (yellow crosses in **b**, x-axis) versus within-sample estimates obtained from LMM-Renv based on test set individuals in the top and bottom 5% strata (not accounting for GxE, using the genotypes and phenotypes of the test set individuals in the strata). Different GIANT variants are coded in colour. **(d)** Analogous scatter plot as in **c**, however displaying the differences between genetic effects in the 5% strata and population estimates of persistent effects (effect sizes due to GxE).



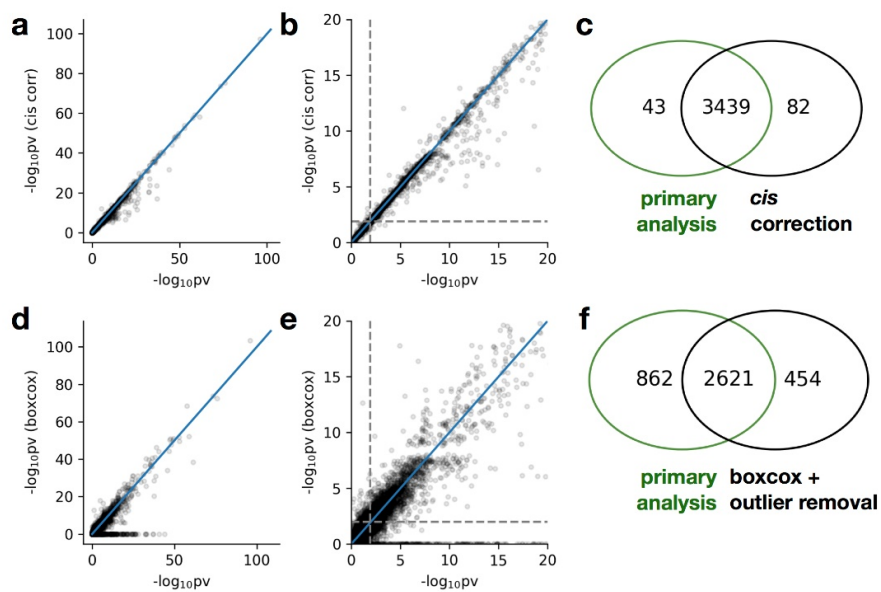


**Supplementary Figure 19 | Rank correlation of per-individual genetic effect sizes across loci for UK Biobank data.** Shown are squared Spearman correlation coefficients of per-individual ( $n = 252,188$ ) allelic effects estimates for 11 GIANT variants with evidence for GxE (more lenient Benjamini-Hochberg adjustment,  $FDR < 0.05$ , **Supp. Table 3**).



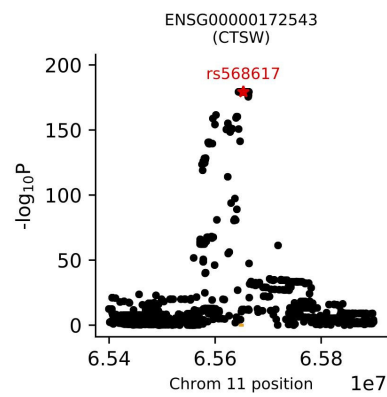
**Supplementary Figure 20 | Statistical calibration and power of alternative interaction tests applied to the blood eQTL dataset.** Compared were the StructLMM interaction test (StructLMM-int) and multi-environment fixed effect tests with as many degrees of freedom as environmental variables (MultiEnv-Renv-LRT-int, labelled MultiEnv-int in the figure), as well as a multi-environment fixed effect test based on 20 principal components of the environmental variables (MultiEnv-20PC-Renv-LRT-int, labelled MultiEnv-20PC-int). Shown are results from applying all methods to test for cell-context interactions at 23,506 lead eQTL variants ( $n = 2,040$ ; **Methods**). **(a)** QQ plot of negative log P values obtained on permuted genotype data. Whereas P values from StructLMM-int and MultiEnv-20PC-Renv-LRT-int were calibrated, MultiEnv-Renv-LRT-int yielded inflated P values, most likely owing to the large number of degrees of the freedom, and hence this model was not considered further. **(b)** Number of interaction eQTL discovered by StructLMM-int and MultiEnv-20PC-Renv-LRT-int as a function of the FDR threshold (Benjamini-Hochberg adjustment). **(c,d)** Scatter plot of negative log P values from StructLMM-int versus MultiEnv-20PC-Renv-LRT-int with a zoom in-view shown in panel **d**. Dashed orange lines correspond to the 5% FDR thresholds. **(e,f)** Overlap of the sets of interaction eQTL identified by StructLMM-int and MultiEnv-20PC-Renv-LRT-int **(e, FDR<5%**; considering all 23,506 tested variants) and interaction eQTL identified in the primary analysis of the data **(f, Zhernakova *et al.*, FDR<5%**; considering 17,952 shared lead eQTL between studies).





**Supplementary Figure 21 | Assessment of the robustness of interaction effects on the blood eQTL data to data pre-processing.** Compared are the StructLMM-int results from the primary analysis and analogous results obtained using alternative data processing strategies ( $n = 2,040$ ). First, to assess the potential of spurious associations due to strong gene-exposure associations, we considered a GxE analysis where for each analysed gene, the additive effect of the lead genetic variant was regressed out from all proxy genes used as environments prior to interaction testing (environments were rank-inverse transformed a second time; **Methods**). **(a)** Scatter plot of the negative log P values from the original (x-axis) and the analysis with environments that were adjusted for *cis* genetic effects (cis corr, y-axis). **(b)** Zoom-in view and **(c)** venn diagram of significant interaction effects (FDR<5%). **(d-f)** Analogous results considering boxcox normalisation followed by removal of outlying samples (gene expression levels that exceeded 2.5 standard deviations) as an alternative pre-processing strategy. **(d-f)** Analogous scatter plots and analysis of overlap as shown in **a-c**. Both control analyses recovered a large fraction of the primary interaction eQTL (98.8% and 75.2% respectively), indicating that the interactions identified are sufficiently robust.





**Supplementary Figure 23 | eQTL association for CTSW.** Manhattan plot of negative log P values obtained from *cis* eQTL mapping of CTSW ( $n = 2,040$ ). The red star indicates the lead eQTL variant, annotated with the identifier of the colocating GWAS variant (*rs568617*,  $r^2=1.00$  associated to Crohn's disease); Manhattan plots for other putative colocalisation events are provided in **Supp. Dataset 1**.

# Supplementary Note

## Contents

<b>1. The StructLMM model</b>	<b>3</b>
1.1. Model definition . . . . .	3
1.2. Definition of the environment covariance . . . . .	4
1.3. Derivation as marginalised linear $G \times E$ interaction model . . . . .	5
1.4. Interaction and joint association tests . . . . .	5
1.4.1. Interaction test . . . . .	5
1.4.2. Association test . . . . .	7
1.4.3. Computational complexity . . . . .	9
1.5. Characterisation of $G \times E$ loci . . . . .	9
1.5.1. Estimation of variance components and the fraction of genetic variance explained by $G \times E$ . . . . .	9
1.5.2. Exploring the environments that drive the observed interaction effects . . . . .	10
1.5.3. Estimation of per-individual allelic effect sizes due to $G \times E$ . . . . .	11
1.5.4. Computational Complexity . . . . .	13
<b>2. Relationship with prior work and comparison methods</b>	<b>13</b>
2.1. Relationship to other $G \times E$ tests . . . . .	13
2.2. Relationship to other LMM implementations . . . . .	15
<b>3. Comparison methods</b>	<b>16</b>
3.1. Single environment models . . . . .	16
3.2. Association tests . . . . .	17
3.2.1. Alternative multivariate $G \times E$ tests based on fixed effects . . . . .	18
3.3. Original analysis of blood eQTL data . . . . .	19
<b>4. Simulations</b>	<b>20</b>
4.1. Environment covariates . . . . .	20
4.2. Phenotype simulation strategy . . . . .	21
4.3. Gene-exposure correlations . . . . .	21
4.4. Skewed and binary environments . . . . .	22

<b>5. Analysis of BMI in UK Biobank</b>	<b>22</b>
5.1. Hold-out validation of per-individual genetic effect sizes . . . . .	22
<b>6. Analysis of cell-context eQTLs in a large blood cohort</b>	<b>23</b>
6.1. Generation of principal components from SAMtools flagstat and Picard tools . . . . .	23
<b>A. Proof for form of Q</b>	<b>23</b>

# 1. The StructLMM model

## 1.1. Model definition

A conventional linear mixed model used for association testing can be cast as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \underbrace{\mathbf{x}\beta_{\text{G}}}_{\text{G}} + \mathbf{u} + \boldsymbol{\psi}, \quad (1)$$

where  $\mathbf{y}$  denotes the  $N \times 1$  phenotype vector for  $N$  individuals,  $\mathbf{x}$  denotes the  $N \times 1$  genotype vector of the focal variant,  $\mathbf{X} \in \mathbb{R}^{N \times K}$  the fixed effect design matrix of  $K$  covariates and  $\mathbf{b} \in \mathbb{R}^{K \times 1}$  their effect sizes. The variable  $\mathbf{u}$  is a random effect and can be used to account for additional additive effects, such as population structure, environment or other additive (confounding) factors and  $\boldsymbol{\psi}$  denotes iid noise. The random effect component  $\mathbf{u}$  and the noise vector  $\boldsymbol{\psi}$  are assumed to follow multivariate normal distributions

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{\Sigma}_{\mathbf{u}}) \quad (2)$$

$$\boldsymbol{\psi} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}_N), \quad (3)$$

where the covariance matrix  $\boldsymbol{\Sigma}_{\mathbf{u}} \in \mathbb{R}^{N \times N}$  is the covariance sample structure of the random effect. Under this conventional linear model, the focal variant  $\mathbf{x}$  is assumed to have a persistent genetic effect on all samples included in the analysis.

StructLMM extends the conventional linear mixed model by including an additional per-individual effect term that accounts for  $\text{G} \times \text{E}$ , which can be represented as an  $N \times 1$  vector,  $\beta_{\text{G} \times \text{E}}$ . The model can be cast as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \underbrace{\mathbf{x}\beta_{\text{G}}}_{\text{G}} + \underbrace{\mathbf{x} \odot \beta_{\text{G} \times \text{E}}}_{\text{G} \times \text{E}} + \underbrace{\mathbf{u}}_{\text{E}} + \boldsymbol{\psi}, \quad (4)$$

where  $\odot$  denotes the element-wise (Hadamard) product. The per-individual effect size vector  $\beta_{\text{G} \times \text{E}}$  is modelled as random effect, following the multivariate normal distribution

$$\beta_{\text{G} \times \text{E}} \sim \mathcal{N}(\mathbf{0}, \sigma_{\text{G} \times \text{E}}^2 \boldsymbol{\Sigma}), \quad (5)$$

where the covariance matrix  $\boldsymbol{\Sigma} \in \mathbb{R}^{N \times N}$  parameterises how per-individual effects covary across individuals and is calculated as a function of observed environmental variables  $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}(\mathbf{E}) \in \mathbb{R}^{N \times N}$ , where  $\mathbf{E}$  is the  $N \times L$  matrix of  $L$  observed environments.

Note that for the special case  $\sigma_{\text{G} \times \text{E}}^2 = 0$ , this model reduces to a standard linear mixed model for genetic association testing. In StructLMM, the random effect component  $\mathbf{u}$  is used to account for additive environmental effects. While in general different covariance functions could be considered for additive environmental effects and interactions, we assume  $\boldsymbol{\Sigma}_{\mathbf{u}} \equiv \boldsymbol{\Sigma}$  for simplicity. Specific choices for the environmental covariance  $\boldsymbol{\Sigma}(\mathbf{E})$  are discussed in Section 1.2.

**Marginal Likelihood** For parameter inference, we consider the marginalised form of the model in Eq (4), which is obtained by integrating over the  $G \times E$  effects  $\beta_{G \times E}$  and the random effect component  $\mathbf{u}$

$$\mathbf{y} \sim \mathcal{N} \left( \mathbf{X}\mathbf{b} + \underbrace{\mathbf{x}\beta_G}_G, \underbrace{\sigma_{G \times E}^2 \text{diag}(\mathbf{x})\Sigma \text{diag}(\mathbf{x})}_{G \times E} + \underbrace{\sigma_e^2 \Sigma}_E + \underbrace{\sigma_n^2 \mathbf{I}}_{\text{noise}} \right). \quad (6)$$

Here, we used the identity  $\mathbf{x} \odot \beta_{G \times E} = \text{diag}(\mathbf{x})\beta_{G \times E}$  in Eq (4), with  $\text{diag}(\mathbf{x})$  denoting the  $N \times N$  diagonal matrix whose diagonal is  $\mathbf{x}$ .

## 1.2. Definition of the environment covariance

In principle, StructLMM can be applied with any valid covariance function [1] of the observed environments. In this work, we only consider linear covariance functions, defined on a potentially large number of continuous and discrete environmental variables

$$\Sigma(\mathbf{E}) = \mathbf{E}\mathbf{E}^T. \quad (7)$$

The use of a linear covariance function was primarily motivated by two appealing properties. First, as the number of samples typically exceeds the number of environments in larger cohorts ( $L \ll N$ ), the linear covariance will be low-rank, enabling parameter inference with a computational complexity that scales linearly with the cohort size (Section 1.4.3). Second, a linear covariance is directly interpretable as there is a one-to-one correspondence between StructLMM and multivariate linear regression using  $L$  covariates to account for the interaction term (see Section 1.3). Note that, depending on the choice of  $\mathbf{E}$ , the linear covariance can be used to (i) model group-specific effects, which correspond to the setting where  $\Sigma$  is a block diagonal matrix with blocks corresponding to the different groups, or per-individual effect sizes (see **Fig. 1b-c**) and (ii) account for non-linear relationships between the observed environmental variables by combining simple observed environments to create more complex ones (similar to the use of basis functions; see below for more details).

**Normalisation of environmental variables** A standard strategy to normalise the variables for defining a linear covariance is to consider standardised features rescaled by  $\frac{1}{\sqrt{L}}$ . This normalisation ensures that the resulting random effect has sample mean 0 and sample variance 1 [2]. Other normalisation procedures, including row standardisation, such that the resulting random effect has per-individual variance 1 or is a correlation matrix, are also valid [1]. Different normalisation procedures correspond to slightly different assumptions regarding the variance explained by different environments within the interaction and additive environment term, similar to building linear covariances using genetic factors [3].

**Accounting for non-linear relationships between environmental variables** Non-linear effects can be accounted for by introducing additional transformed environmental variables, similar to using basis functions for non-linear regression. For example,

to model possible dependencies of  $G \times E$  and additive environmental effects of  $L$  environments  $\mathbf{e}_1, \dots, \mathbf{e}_L$  with a categorical variable  $\mathbf{c} \in \{0, 1\}^{N \times 1}$  (e.g. gender), one can define  $\Sigma$  using the extended set of environments  $\mathbf{E} = [\mathbf{c} \otimes \mathbf{e}_1, \dots, \mathbf{c} \otimes \mathbf{e}_L, (\mathbf{1}_N - \mathbf{c}) \otimes \mathbf{e}_1, \dots, (\mathbf{1}_N - \mathbf{c}) \otimes \mathbf{e}_L] \in \mathbb{R}^{N \times 2L}$ . This approach is used to define a gender-adjusted and age-adjusted environmental covariance.

### 1.3. Derivation as marginalised linear $G \times E$ interaction model

When using a linear environmental covariance, StructLMM can be interpreted as a random-effect implementation of the multi-environment linear interaction model where persistent genetic and  $G \times E$  effects are modelled as fixed effects. This relationship is similar to the well-known equivalence between multivariate Bayesian linear regression and a linear mixed model (e.g. [4]). Briefly, denoting with  $\mathbf{e}_1, \dots, \mathbf{e}_L$  the set of observed  $N \times 1$  vectors for  $L$  environmental variables, the linear model underlying StructLMM can be cast as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{\sum_{l=1}^L (\mathbf{x} \odot \mathbf{e}_l)\beta_l}_{\text{G} \times \text{E}} + \underbrace{\sum_{l=1}^L \mathbf{e}_l\alpha_l}_{\text{E}} + \boldsymbol{\psi}, \quad (8)$$

where  $\boldsymbol{\psi} \sim \mathcal{N}(\mathbf{0}, \sigma_n^2 \mathbf{I}_N)$ . Defining prior variances on  $G \times E$  and additive environmental effects

$$\beta_l \sim \mathcal{N}(0, \sigma_{G \times E}^2), \quad (9)$$

$$\alpha_l \sim \mathcal{N}(0, \sigma_e^2), \quad (10)$$

and marginalising over  $\beta_l$  and  $\alpha_l$  results in the marginalised StructLMM model in Eq. (6). It is of note that both the association and interaction tests in StructLMM can also be directly implemented in the fixed effect framework, without marginalization. These alternative tests are considered in **Supp. Fig. 1b**, where we observe that fixed effect tests are not always calibrated and are underpowered in some settings. See Section 3.2.1 for full details on the implementation of alternative methods that are compared to StructLMM.

### 1.4. Interaction and joint association tests

We define two variance component score tests based on the StructLMM model, (i) an interaction test to identify loci with  $G \times E$  and (ii) an association test that accounts for heterogeneity in genetic effect sizes due to  $G \times E$  (jointly testing the  $G$  and  $G \times E$  effects).

#### 1.4.1. Interaction test

Using the marginalised model in Eq. 6, a test for  $G \times E$  interactions corresponds to the alternative hypothesis  $\sigma_{G \times E}^2 > 0$ . We define an efficient score-based test that enables the calculation of P values with a complexity that scales linearly in the number of



individuals, provided that the environment covariance  $\Sigma$  is low rank. We note that the null of the interaction test reduces to a standard linear mixed model with a low-rank covariance matrix for additive random effects, for which existing efficient inference strategies can be reused [5].

**Score test for interaction test** In the model in Eq (4-5), the score-test statistics can be computed analogously to the procedure described in [6]

$$\begin{aligned}
Q &= \frac{1}{2} \mathbf{y}^T \mathbf{P} \mathbf{K}_1 \mathbf{P} \mathbf{y} \\
&= \frac{1}{2} \mathbf{y}^T \mathbf{P} (\text{diag}(\mathbf{x}) \Sigma \text{diag}(\mathbf{x})) \mathbf{P} \mathbf{y} \\
&= \frac{1}{2} \mathbf{y}^T \mathbf{P} \underbrace{[\text{diag}(\mathbf{x}) \mathbf{E}]}_{\mathbf{W}} [\text{diag}(\mathbf{x}) \mathbf{E}]^T \mathbf{P} \mathbf{y} \\
&= \frac{1}{2} \|\mathbf{W}^T \mathbf{P} \mathbf{y}\|^2.
\end{aligned} \tag{11}$$

where we have defined

$$\begin{aligned}
\mathbf{K}_1 &= \text{diag}(\mathbf{x}) \Sigma \text{diag}(\mathbf{x}), \\
\mathbf{W} &= \text{diag}(\mathbf{x}) \mathbf{E}, \\
\mathbf{P} &= \mathbf{H}_0^{-1} - \mathbf{H}_0^{-1} [\mathbf{X}, \mathbf{x}] ([\mathbf{X}, \mathbf{x}]^T \mathbf{H}_0^{-1} [\mathbf{X}, \mathbf{x}])^{-1} [\mathbf{X}, \mathbf{x}]^T \mathbf{H}_0^{-1}.
\end{aligned} \tag{12}$$

The matrix  $\mathbf{H}_0$  denotes the total covariance matrix estimated under the null model

$$\mathbf{H}_0 = \hat{\sigma}_e^2 \Sigma + \hat{\sigma}_n^2 \mathbf{I}, \tag{13}$$

where  $\hat{\sigma}_e^2$  and  $\hat{\sigma}_n^2$  correspond to the (null model) maximum likelihood estimates (MLE) of  $\sigma_e^2$  and  $\sigma_n^2$ . It can be shown that  $Q$  follows a mixture of  $\chi^2$  distributions [6, 7]

$$Q \sim \sum_k a_k \chi_1^2, \tag{14}$$

where the vector of the coefficients  $\mathbf{a} = [a_k]_k$  can be computed as the eigenvalues of  $\mathbf{P}^{\frac{T}{2}} \mathbf{K}_1 \mathbf{P}^{\frac{1}{2}}$

$$\mathbf{a} = \text{eigh} \left( \mathbf{P}^{\frac{T}{2}} \mathbf{K}_1 \mathbf{P}^{\frac{1}{2}} \right) \tag{15}$$

$$= \text{eigh} (\mathbf{W}^T \mathbf{P} \mathbf{W}). \tag{16}$$

As the distribution of the score test statistics is a mixture of  $\chi^2$ , following the procedure in SKAT [6], P values are computed using Davies method [8] (an exact method which directly inverts the characteristic function), switching to the Liu method [9] (an approximation method based on moment matching) when the Davies method fails to converge (see [10] for full details).

### 1.4.2. Association test

To define a joint association test that accounts for the possibility of effect size heterogeneity, we consider a fully marginalised form of the model described in Eq. 4, where both the  $G \times E$  and the persistent effect effect sizes, following distributions  $\beta_{G \times E} \sim \mathcal{N}(\mathbf{0}, \sigma_{G \times E}^2 \Sigma)$  and  $\beta_G \sim \mathcal{N}(0, \sigma_G^2)$ , respectively, are integrated out. This fully marginalised model can be cast as

$$\mathbf{y} \sim \mathcal{N} \left( \mathbf{X}\mathbf{b}, \underbrace{\sigma_{G_{\text{tot}}}^2 \mathbf{K}_\rho}_{G+G \times E} + \underbrace{\sigma_e^2 \Sigma}_{E} + \underbrace{\sigma_n^2 \mathbf{I}}_{\text{noise}} \right) \quad (17)$$

where  $\sigma_{G_{\text{tot}}}^2 = \sigma_G^2 + \sigma_{G \times E}^2$  denotes the total variance of the genetic effect (including both persistent and  $G \times E$  components). Additionally, we have introduced

$$\mathbf{K}_\rho = \underbrace{(1 - \rho) \mathbf{x}\mathbf{x}^T}_G + \underbrace{\rho \text{diag}(\mathbf{x}) \Sigma \text{diag}(\mathbf{x})}_{G \times E}, \quad (18)$$

where the parameter  $\rho$  is defined as the fraction of the total genetic variance explained by  $G \times E$  effects,  $\rho = \sigma_{G \times E}^2 / \sigma_{G_{\text{tot}}}^2$ . We note that this fully marginalised form of the model can also be derived from the linear model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{x} \odot \beta + \mathbf{u} + \psi, \quad (19)$$

by marginalizing over  $\beta$ , assuming the following multivariate normal prior

$$\beta \sim \mathcal{N} \left( \mathbf{0}, \sigma_{G_{\text{tot}}}^2 \left[ \underbrace{(1 - \rho) \mathbf{1}\mathbf{1}^T}_G + \underbrace{\rho \Sigma}_{G \times E} \right] \right). \quad (20)$$

Note that in this representation  $\beta$  accounts for  $G$  and  $G \times E$  effects, with the first covariance term corresponding to a persistent genetic effect ( $G$ , full correlation of the genetic effect sizes across individuals) whereas the second covariance term accounts for heterogeneity of genetic effect sizes due to  $G \times E$ .

In the marginalised form of the model given by Eq (17-18), an association test corresponds to assessing the alternative hypothesis  $\sigma_{G_{\text{tot}}}^2 > 0$ . The form of both the model and the test allow us to implement an efficient optimal score test procedure, building on the work in [11]. This score-based test again has complexity that scales linear in the number of individuals, provided that the environment covariance  $\Sigma$  is low rank. Below we present details of how the test is implemented, first considering the case of known  $\rho$  followed by general case of unknown  $\rho$ .

**Score test for given  $\rho$ .** The score test for given  $\rho$  is analogous to the one described in Section 1.4.1 but with  $\mathbf{K}_1$  replaced by  $\mathbf{K}_\rho = ((1 - \rho) \mathbf{x}\mathbf{x}^T + \rho \text{diag}(\mathbf{x}) \Sigma \text{diag}(\mathbf{x}))$  such that  $\mathbf{W} = [\sqrt{1 - \rho} \mathbf{x} \quad \sqrt{\rho} \text{diag}(\mathbf{x}) \mathbf{E}]$  and  $[\mathbf{X}, \mathbf{x}]$  replaced by  $\mathbf{X}$ .

**Score test for unknown  $\rho$ .** As in the test  $\sigma_{G_{\text{tot}}}^2 > 0$  in the marginalised model in Eq (17-18), the fraction of genetic variance due to  $G \times E$  ( $\rho$ ) is in general unknown, one would like to select  $\rho$  to maximise detection power. Following [11], we define the optimal score test as follows:

1. Define a grid of  $R$  values  $\rho_1 < \rho_2 < \dots < \rho_R$  for  $\rho_r$  (by default, we consider  $[0, 0.5, 0.75, 0.84, 0.91, 0.96, 0.99, 1]$ );
2. For each value  $\rho_r$ , compute the score test statistic  $Q_{\rho_r}$  using Eq (11) and corresponding P values  $p_{\rho_r}$  using the modified Liu method (an approximation method);
3. Compute a P value for the test statistic  $T = \min \{p_{\rho_1}, \dots, p_{\rho_R}\}$ .

By definition, this P value can be computed from  $Q_{\rho_1}, \dots, Q_{\rho_R}$  and  $T$  as

$$P_T = p(t < T) = 1 - P(Q_{\rho_1} < q_{\rho_1}(T), Q_{\rho_2} < q_{\rho_2}(T), \dots, Q_{\rho_R} < q_{\rho_R}(T)) \quad (21)$$

where  $q_{\rho_r}(T)$  denotes the  $(1 - T)^{\text{th}}$  percentile of  $Q_{\rho_r}$ . It can be shown that  $Q_{\rho}$  can be written as follows (see derivation in Appendix A)

$$Q_{\rho} = \frac{1}{2}\rho\kappa + \frac{1}{2}\tau_{\rho}\eta_0, \quad (22)$$

where

$$\kappa = \sum_{k=1}^m \lambda_k \eta_k + \xi \quad (23)$$

$$\boldsymbol{\lambda} = \text{eigh}(\boldsymbol{\Lambda}) \quad (24)$$

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_0 - \frac{1}{c}\boldsymbol{\alpha}\boldsymbol{\alpha}^T \quad (25)$$

$$\text{Var}(\xi) = \frac{4}{c}\boldsymbol{\alpha}^T \boldsymbol{\Lambda}_0 \boldsymbol{\alpha} - \frac{4}{c^2}(\boldsymbol{\alpha}^T \boldsymbol{\alpha})^2 \quad (26)$$

$$\tau_{\rho} = c(1 - \rho) + \frac{\rho}{c}\boldsymbol{\alpha}^T \boldsymbol{\alpha} \quad (27)$$

$$\eta_k \stackrel{\text{iid}}{\sim} \chi_1^2 \quad (28)$$

$$\boldsymbol{\Lambda}_0 = \boldsymbol{E}^T \text{diag}(\boldsymbol{x}) \boldsymbol{P} \text{diag}(\boldsymbol{x}) \boldsymbol{E} \quad (29)$$

$$\boldsymbol{\alpha} = \boldsymbol{E}^T \text{diag}(\boldsymbol{x}) \boldsymbol{P} \boldsymbol{x} \quad (30)$$

$$c = \boldsymbol{x} \boldsymbol{P} \boldsymbol{x}^T. \quad (31)$$

Replacing Eq (22) in Eq (21) results in

$$P_T = p(t < T) = 1 - \mathbb{E} \left[ P(\kappa < \min \{ (2q_{\rho_r}(T) - \tau_{\rho_r}\eta_0) / \rho_r \}_{r=1}^R) \right], \quad (32)$$

This P value can be computed using one-dimensional numerical integration in the same manner as for the interaction test; using Davies method [8] (an exact method which directly inverts the characteristic function), switching to the Liu method [9] (an approximation method based on moment matching) when convergence fails (see [11, 10] for details).

### 1.4.3. Computational complexity

To evaluate the interaction and joint association tests, we need to first fit the null model and then compute the corresponding score test statistics. A naive implementation of both of these operations would result in computations that scale cubically with the number of individuals, i.e.  $O(N^3)$ . To reduce the complexity of both operations, we exploit that, when considering a linear covariance of  $L$  environments, the total covariance matrix is

$$\sigma_e^2 \mathbf{E}\mathbf{E}^T + \sigma_n^2 \mathbf{I}, \quad (33)$$

where the first term is low rank ( $L \ll N$ ). In this setting, the null model can be fit with computational complexity  $O(NR^2 + R^3)$ , where  $R$  is the rank of the first covariance term (in Eq (33), in this case,  $R = L$ ). We refer to [7, 12] for further details.

Concerning the computation of the score test statistics for interaction test,  $Q$  can be calculated with computational complexity  $O(NL^2 + NK^2 + NLK + L^2K + K^3 + L^3)$  where  $K$  corresponds to the number of covariates, while the coefficient  $\mathbf{a}$  of the  $\chi_1^2$ -mixture (eigenvalues of  $\mathbf{W}^T \mathbf{P} \mathbf{W}$ , see Eq (16)), can be computed in  $O(NL^2 + NK^2 + NLK + L^2K + K^3 + L^3)$ . The computation of P values either using the Davies or Liu method does not depend on the number of individuals.

For the association test for unknown  $\rho$ , the standard score-based procedure is repeated multiple times, corresponding to the number of values in the grid search over  $\rho$ . The additional variables that need to be computed (see Eqs 23-31) have linear computational complexity with the number of individuals. Finally, the one-dimensional numerical integration step (using Davies or Liu method) has a computational complexity that is independent of the number of individuals.

## 1.5. Characterisation of G×E loci

The StructLMM framework facilitates different analyses for characterising G×E effects. Although all the computations of these analyses remain linear in the number of individuals, the underlying computations are less efficient than those required for score tests, as maximum likelihood model parameters need to be determined explicitly for each variant that is considered. For this reason, we recommend using these analyses exclusively at the suggestive loci identified by StructLMM association and interaction tests.

### 1.5.1. Estimation of variance components and the fraction of genetic variance explained by G×E

StructLMM allows for decomposing the phenotypic variance into genetic (G), G×E and additive environment (E) variance components. Based on the marginalised model

in Eq (6), the three variance components can be estimated as

$$\text{var}^{(G)} = \text{var}_S(\mathbf{x}\hat{\beta}_G) = \hat{\beta}_G^2 \text{var}_S(\mathbf{x}), \quad (34)$$

$$\text{var}^{(G \times E)} = \text{var}_M(\hat{\sigma}_{G \times E}^2(\mathbf{x} \odot \mathbf{E})(\mathbf{x} \odot \mathbf{E})^T), \quad (35)$$

$$\text{var}^{(E)} = \text{var}_M(\hat{\sigma}_E^2 \mathbf{E} \mathbf{E}^T). \quad (36)$$

Here,  $\text{var}_S$  denotes the sample variance and  $\text{var}_M(\mathbf{K})$  denotes the expected value of the sample variance of  $\mathbf{z}$  with  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ . We also use the following two identities for terms that involve the linear covariance matrix,  $\mathbf{\Sigma} = \mathbf{E} \mathbf{E}^T$ ,  $\text{diag}(\mathbf{x}) \mathbf{\Sigma} \text{diag}(\mathbf{x}) = (\mathbf{x} \odot \mathbf{E})(\mathbf{x} \odot \mathbf{E})^T$  and denote the MLE estimates of  $\beta_G$ ,  $\sigma_{G \times E}^2$  and  $\sigma_E^2$  as  $\hat{\beta}_G$ ,  $\hat{\sigma}_{G \times E}^2$  and  $\hat{\sigma}_E^2$ , respectively.

Using  $\text{var}_M(\mathbf{K}) = \frac{1}{N-1} \text{tr}(\mathbf{P} \mathbf{K})$ , where  $\mathbf{P} = \mathbf{I} - \frac{1}{\text{dim}(\mathbf{K})} \mathbf{1} \mathbf{1}^T$  [2, 13], we obtain

$$\text{var}^{(G \times E)} = \frac{1}{N-1} \hat{\sigma}_{G \times E}^2 \|\mathbf{P}(\mathbf{x} \odot \mathbf{E})\|^2 \quad (37)$$

$$\text{var}^{(E)} = \frac{1}{N-1} \hat{\sigma}_E^2 \|\mathbf{P} \mathbf{E}\|^2. \quad (38)$$

Finally, the estimated fraction of genetic variance explained by  $G \times E$  effects ( $\rho$ ) follows as

$$\hat{\rho} = \frac{\text{var}^{(G \times E)}}{\text{var}^{(G)} + \text{var}^{(G \times E)}}. \quad (39)$$

It can be seen that this is equivalent to  $\rho$  described in Section 1.4.2 (optimal score test procedure), such that  $\text{var}^{(G \times E)} \equiv \sigma_{G \times E}^2$  and  $\text{var}^{(G)} + \text{var}^{(G \times E)} \equiv \sigma_{\text{tot}}^2$ .

Note that the estimate  $\hat{\rho}$  is based on a MLE whilst the estimate obtained from the optimal score test procedure corresponds to the minimum P value obtained from searching over a grid of predefined values, which has a different objective.

### 1.5.2. Exploring the environments that drive the observed interaction effects

The evidence of individual environmental variables or sets of environments for driving the observed  $G \times E$  effects can be assessed by comparing the log marginal likelihood of models with and without these environments included. The resulting Bayes factors from such comparisons are directly calibrated because the number of parameters fit using maximum likelihood is independent of the number of environmental variables.

Given a variant and a set of  $L$  environments  $\mathcal{E} = \{\mathbf{e}_1, \dots, \mathbf{e}_L\}$ , the evidence for a subset of environments,  $\mathcal{E}_i \subseteq \mathcal{E}$ ,  $L_i < L$  driving the observed  $G \times E$  effect can be estimated as

$$\log(\text{Bayes factor})(\mathcal{E}_i) = \text{LML}(\mathbf{M}_{\mathcal{E}}) - \text{LML}(\mathbf{M}_{\mathcal{E} \setminus \mathcal{E}_i}) \quad (40)$$

---

\*We use the  $\odot$  operator between vectors and matrices to denote the matrix obtained by the element-wise multiplication of matrix columns by the input vector.

where  $\text{LML}(\mathbf{M}_{\mathcal{E}})$  and  $\text{LML}(\mathbf{M}_{\mathcal{E} \setminus \mathcal{E}_i})$  denotes the marginal log-likelihood of the model in Eq (6), either considering the full or reduced sets of environments to define the  $\mathbf{G} \times \mathbf{E}$  environment covariance respectively. Specifically, while in  $\mathbf{M}_{\mathcal{E}}$  all environments are used in the  $\mathbf{G} \times \mathbf{E}$  covariance, in  $\mathbf{M}_{\mathcal{E} \setminus \mathcal{E}_i}$  only the environments not in  $\mathcal{E}_i$  are considered (the operation  $\setminus$  denotes the set difference). Note that the additive environment covariance remains the same in both models and is always estimated using all environments. To identify a putative causal set of driving environments, we use a greedy backward elimination procedure. Initially, we calculate the log(Bayes factor) for each environmental variable, to identify the marginal environment with the most evidence:  $i_{\max} = \text{argmax}_i (\log(\text{Bayes factor})(\mathcal{E}_i))$ . Subsequently, this environment is removed and the process iterated, stopping when there is positive evidence based on the log Kass and Raftery scale [14] that we have selected a full set of environments that can drive the observed  $\mathbf{G} \times \mathbf{E}$  effect (i.e. when the delta log(Bayes Factor) between the model under consideration and the model that removes all environmental variables is  $< 1$ ).

### 1.5.3. Estimation of per-individual allelic effect sizes due to $\mathbf{G} \times \mathbf{E}$

StructLMM allows for estimating per-individual allelic effect sizes of selected variants based on the distribution of environmental profiles attained in a population. For a given locus, the effect size of the non-reference allele in environment profile  $\mathbf{e}^*$  can be estimated using a procedure based on the best linear unbiased predictor (BLUP, [15]):

1. Estimate the  $\mathbf{G} + \mathbf{G} \times \mathbf{E}$  effect for a homozygous reference individual in environmental profile  $\mathbf{e}^*$  using BLUP and the model in Eq (6)

$$\text{ref}(\mathbf{e}^*) = \underbrace{x_r \hat{\beta}_G}_{\mathbf{G}} + \underbrace{\hat{\sigma}_{\mathbf{G} \times \mathbf{E}}^2 x_r \mathbf{e}^{*T} (\mathbf{x} \odot \mathbf{E})^T \hat{\mathbf{H}}^{-1} (\mathbf{y} - \mathbf{X} \hat{\mathbf{b}} - \mathbf{x} \hat{\beta}_G)}_{\mathbf{G} \times \mathbf{E}}. \quad (41)$$

Here,  $\hat{\mathbf{H}}$ ,  $\hat{\beta}$  and  $\hat{\mathbf{b}}$  are MLE of the total covariance matrix, the genetic effect size  $\beta$  and the covariate effect sizes  $\mathbf{b}$  in the model in Eq (6), while  $x_r$  is the value corresponding to the encoding of the homozygous reference genotype.

2. Compute the BLUP of the  $\mathbf{G} + \mathbf{G} \times \mathbf{E}$  for a heterozygous individual in environmental profile  $\mathbf{e}^*$  using the same model

$$\text{alt}(\mathbf{e}^*) = \underbrace{x_a \hat{\beta}_G}_{\mathbf{G}} + \underbrace{\hat{\sigma}_{\mathbf{G} \times \mathbf{E}}^2 x_a \mathbf{e}^{*T} (\mathbf{x} \odot \mathbf{E})^T \hat{\mathbf{H}}^{-1} (\mathbf{y} - \mathbf{X} \hat{\mathbf{b}} - \mathbf{x} \hat{\beta}_G)}_{\mathbf{G} \times \mathbf{E}}, \quad (42)$$

where  $x_a$  corresponds to the encoding of the heterozygous genotype.

3. The allelic effect in environment profile  $\mathbf{e}^*$  is then defined as the difference be-

tween  $\text{alt}(\mathbf{e}^*)$  and  $\text{ref}(\mathbf{e}^*)$

$$\beta^{(\text{tot})}(\mathbf{e}^*) = \text{alt}(\mathbf{e}^*) - \text{ref}(\mathbf{e}^*) \quad (43)$$

$$= (x_a - x_r) \left( \underbrace{\hat{\beta}_G}_G + \underbrace{\hat{\sigma}_{G \times E}^2 \mathbf{e}^{*T} (\mathbf{x} \odot \mathbf{E})^T \hat{\mathbf{H}}^{-1} (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{x}\hat{\beta}_G)}_{G \times E} \right) \quad (44)$$

$$= (x_a - x_r) \left( \underbrace{\hat{\beta}_G}_G + \underbrace{\beta_{G \times E}(\mathbf{e}^*)}_{G \times E} \right) \quad (45)$$

In order to compute allelic effects for  $N^*$  different environmental profiles  $\{\mathbf{e}_1^*, \dots, \mathbf{e}_{N^*}^*\}$ , we use Eq (44) in vectorial form as

$$\boldsymbol{\beta}^{(\text{tot})}(\mathbf{E}^*) = (x_a - x_r) \left( \underbrace{\mathbf{1}_N \hat{\beta}_G}_G + \underbrace{\hat{\sigma}_{G \times E}^2 \mathbf{E}^{*T} (\mathbf{x} \odot \mathbf{E})^T \hat{\mathbf{H}}^{-1} (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{x}\hat{\beta}_G)}_{G \times E} \right), \quad (46)$$

where  $\boldsymbol{\beta}^{(\text{tot})}$  is  $N^* \times 1$  vector of genetic effects in the corresponding  $N^*$  profiles and  $\mathbf{E}^*$  is the  $N^* \times L$  environment matrix

$$\mathbf{E}^* = \begin{bmatrix} \mathbf{e}_1^* \\ \dots \\ \mathbf{e}_{N^*}^* \end{bmatrix}. \quad (47)$$

This formula can be applied for the whole set of  $N$  observed environmental profiles ( $\mathbf{E}^* = \mathbf{E}$ ) (in-sample estimations), for environmental profiles from test/validation samples (out-of-sample predictions).

Finally, we note that if the genetic vector  $\mathbf{x}$  is standardised, then  $x_r = \frac{-2p}{\sqrt{2p(1-p)}}$  and  $x_a = \frac{1-2p}{\sqrt{2p(1-p)}}$ , where  $p$  is the observed minor allele frequency (MAF). It follows that  $x_a - x_r = \frac{1}{\sqrt{2p(1-p)}}$ .

**Aggregate environment driving  $G \times E$ .** StructLMM can also be used to estimate the aggregate environment to explain  $G \times E$  at single loci, which can be interpreted as the MAP posterior estimate of  $\beta_{G \times E}$ . An estimate can be derived based form the linear model equivalence underlying StructLMM in Eq (8) (see Section 1.3), which we use to rewrite the  $G \times E$  term as

$$\sum_{l=1}^L (\mathbf{x} \odot \mathbf{e}_l) \beta_l = \mathbf{x} \odot \left( \sum_{l=1}^L \mathbf{e}_l \beta_l \right) = \mathbf{x} \odot (\mathbf{E} \boldsymbol{\beta}'_{G \times E}), \quad (48)$$

where  $\boldsymbol{\beta}'_{G \times E} = [\beta_1, \beta_2, \dots, \beta_L]^T$  and  $\mathbf{E} \boldsymbol{\beta}'_{G \times E}$  denotes the aggregate environment driving  $G \times E$ . Using the StructLMM model in Eq (6), a maximum a posteriori estimate of the

aggregate environment is  $\mathbf{E}\hat{\boldsymbol{\beta}}'_{\mathbf{G} \times \mathbf{E}}$ , where

$$\hat{\boldsymbol{\beta}}'_{\mathbf{G} \times \mathbf{E}} = \hat{\sigma}_{\mathbf{G} \times \mathbf{E}}^2 (\mathbf{x} \odot \mathbf{E})^T \hat{\mathbf{H}}^{-1} (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{x}\hat{\beta}_{\mathbf{G}}). \quad (49)$$

#### 1.5.4. Computational Complexity

The post-processing analyses for characterising  $\mathbf{G} \times \mathbf{E}$  effects build on the marginal model in Eq. (6), which is amenable to fast inference schemes. Specifically, introducing  $\mathbf{Z} = [\sigma_{\mathbf{G} \times \mathbf{E}} \text{diag}(\mathbf{x})\mathbf{E}, \sigma_{\mathbf{E}}\mathbf{E}]$ , the model in Eq (6) can be written as

$$\mathbf{y} \sim \mathcal{N} \left( \begin{bmatrix} \mathbf{X} & \mathbf{x} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \beta \end{bmatrix}, \mathbf{Z}\mathbf{Z}^T + \sigma_n^2 \mathbf{I} \right). \quad (50)$$

Note, that this model has the same form as the null model, where the first covariance term has rank  $2L$ . Maximum likelihood can thus be achieved with computational complexity  $O(N(2L)^2 + (2L)^3)$  (see [7, 12] for details).

## 2. Relationship with prior work and comparison methods

In this section, we review existing methods for  $\mathbf{G} \times \mathbf{E}$  and we describe LMMs that are otherwise related to StructLMM. In Section 2.1 we review related existing  $\mathbf{G} \times \mathbf{E}$  testing strategies; in Section 2.2 we discuss technical similarities between StructLMM and existing LMM-based approaches. Finally, in Section 3 we describe alternative methods that are compared with StructLMM.

### 2.1. Relationship to other $\mathbf{G} \times \mathbf{E}$ tests

**Single-environment  $\mathbf{G} \times \mathbf{E}$  tests** A standard linear model to test for  $\mathbf{G} \times \mathbf{E}$  using quantitative traits and a single environmental variable can be cast as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{x}\beta_{\mathbf{G}} + (\mathbf{x} \odot \mathbf{e})\beta_{\mathbf{G} \times \mathbf{E}} + \mathbf{e}\beta_{\mathbf{E}} + \boldsymbol{\psi}, \quad (51)$$

where  $\mathbf{y}$  is a  $N \times 1$  vector of quantitative trait measurements for  $N$  individuals,  $\mathbf{X}$  is the  $N \times K$  fixed effect design matrix for  $K$  covariates,  $\mathbf{b}$  is the  $K \times 1$  vector of their effect sizes,  $\mathbf{x}$  is an  $N \times 1$  genotype vector for the variant under consideration,  $\beta_{\mathbf{G}}$  the corresponding persistent genetic effect,  $\mathbf{e}$  an  $N \times 1$  environment vector for the studied environment,  $\beta_{\mathbf{G} \times \mathbf{E}}$  the genotype-environment interaction effect and  $\boldsymbol{\psi}$  a noise term, typically assumed to be iid normal  $\boldsymbol{\psi} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_n^2)$  [16]. The model in Eq (51) can be used both for testing  $\mathbf{G} \times \mathbf{E}$  ( $\beta_{\mathbf{G} \times \mathbf{E}} \neq 0$ ) or to test for association while accounting for  $\mathbf{G} \times \mathbf{E}$  ( $[\beta_{\mathbf{G}}, \beta_{\mathbf{G} \times \mathbf{E}}] \neq \mathbf{0}$ ) [17, 18].

**$\mathbf{G} \times \mathbf{E}$  variant-set tests** The model in Eq (51) can be extended to aggregate effects across multiple variants, enabling  $\mathbf{G} \times \mathbf{E}$  variant-set tests for single environments. For



a set of  $R$  genetic variants  $\{\mathbf{g}_1, \dots, \mathbf{g}_R\}$ , a variant-set model can be cast as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \sum_{s=1}^R \mathbf{g}_s \beta_s^G + \sum_{s=1}^R (\mathbf{g}_s \odot \mathbf{e}) \beta_s^{G \times E} + \mathbf{e} \beta^E + \boldsymbol{\psi}. \quad (52)$$

Tukey’s one degree-of-freedom (df) is one of the first  $G \times E$  variant-set tests [19], which assumes proportionality between interaction and marginal genetic effect sizes (i.e.,  $\beta_s^{G \times E} = \theta \beta_s^G$ ,  $s = 1, \dots, R$ ). This assumption can be restrictive, but comes with the advantage that variant-sets can be tested via  $\theta \neq 0$  (1 df). Other strategies to re-weight the interaction effects and limit the number of testing parameters have been proposed (for example, Jiao et al [20] proposed using gene-environment correlations).

Another reweighting scheme that is commonly used for  $G \times E$  analysis is known as the genetic risk score (GRS), where one assumes  $\beta_s^G = w_s \beta^G$  and  $\beta_s^{G \times E} = w_s \beta^{G \times E}$ . In GRS analyses, the SNP weight  $w_s$  is either i) set to  $w_s = 1$  for all  $s$  (unweighted GRS, [21]) or ii) to  $w_s = \beta_s^P$ , where  $\beta_s^P$  is the effect size of variant  $s$  identified from a previous study that tested only for main genetic effects (weighted GRS, [22]). GRS analyses have been used in  $G \times E$  interaction analyses of BMI in UK Biobank data [22, 23] using a weighted score based on the effect sizes estimated from the 2015 GIANT association analysis for BMI [24].

A class of  $G \times E$  variant-set tests that is technically related to StructLMM is based on random effects for  $\boldsymbol{\beta}^{G \times E}$ , which build on linear mixed models, similarly to StructLMM. In the following we describe the interaction sequence kernel association test (GESAT) [25], as representative of a family of related methods [25, 26, 27, 28]. Starting from equation Eq (52), GESAT can be derived by defining a multivariate normal prior on the genetic effect sizes,  $\boldsymbol{\beta}^{G \times E} \sim \mathcal{N}(\mathbf{0}, \tau \mathbf{I}_R)$ , where interactions can be tested by assessing  $\tau \neq 0$  using a score test (see Section 1.4). GESAT models the additive genetic effects in the null model using fixed effects, estimated using ridge regression to mitigate the large number of covariates in the test [29]. The score-based test statistics of the interaction test is

$$\mathbf{Q} = (\mathbf{y} - \hat{\boldsymbol{\mu}})^\top \mathbf{S} \mathbf{S}^\top (\mathbf{y} - \hat{\boldsymbol{\mu}}), \quad (53)$$

where  $\mathbf{S} = [\mathbf{g}_1 \odot \mathbf{e}, \dots, \mathbf{g}_R \odot \mathbf{e}]$  and  $\hat{\boldsymbol{\mu}}$  is the optimised mean under the null model.  $\mathbf{Q}$  follows a mixture of  $\chi^2$  distributions with 1 df [6] (see Section 1.4). Among the generalisations of GESAT are methods tailored towards rare variants (iSKAT, [26]), with a prior on  $\boldsymbol{\beta}^{G \times E}$  that accounts both for scenarios where the  $G \times E$  effects have consistent effect directions or have independent effects ( $\boldsymbol{\beta}^{G \times E} \sim \mathcal{N}(\mathbf{0}, \tau(\pi \mathbf{1}_R \mathbf{1}_R^\top + (1 - \pi) \mathbf{I}_R))$ ). An optimal score-test is considered to assess statistical significance (see Section 1.4). The testing procedure in iSKAT is related to the association test in StructLMM, but there are important differences. First, the prior used in iSKAT is designed to empower the detection of interaction effects between multiple rare variants and a single environmental variable. Conversely, StructLMM considers a prior on the effect of a single variants due to  $G \times E$  with multiple environments. Additionally, a technical difference is in how the two methods deal with the high number of covariates: while StructLMM

uses a random effect, iSKAT uses ridge regression. Finally, StructLMM can also be used to efficiently fit the full model with  $G+E+G \times E$  effects, which is critical to enable for the characterisation of selected  $G \times E$  loci (see Section 1.5).

StructLMM is also related to a recently proposed random effect set test, iSet [30], which builds on a framework for multi-trait set tests [12]. Critically, the model is limited to analyses of moderately sized cohorts (at most tens of thousands of samples), it only considers categorical environments and it cannot be used for the analysis of multiple environments.

**Existing multi-environment  $G \times E$  analyses** The only multivariate  $G \times E$  analysis with multiple environments we are aware of has been proposed in [31], in which the authors study  $G \times E$  effects on BMI at the *FTO* locus. Briefly, the authors employ a two-step approach, first selecting relevant environmental variables based on their additive effect on the phenotype (using a cross validation scheme), and then assess  $G \times E$  effects using a z-score test statistics using a multivariate fixed effect model (as in Eq. (8)). Thus, the method is similar to the fixed effect model we implement for comparison (Section 1.5.2). At present, there is no software that implements this procedure [31], which also does not scale to genome-wide applications due to the costly cross-validation step.

## 2.2. Relationship to other LMM implementations

Although the models follow different aims, StructLMM is technically related to the optimal score-based test originally implemented in SKAT-O [11], a rare variant association test. The SKAT-O model can be cast as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{G}\boldsymbol{\beta}^G + \boldsymbol{\psi}, \quad (54)$$

where  $\mathbf{G} = [\mathbf{g}_1, \dots, \mathbf{g}_R]$  denote the  $N \times R$  genetic design matrix for  $R$  rare variants and  $\boldsymbol{\beta}^G$  is set to follow the multivariate normal distribution

$$\boldsymbol{\beta}^G \sim \mathcal{N}(\mathbf{0}, \tau(\pi \mathbf{1}_R \mathbf{1}_R^T + (1 - \pi) \mathbf{I}_R)). \quad (55)$$

This formulation interpolates between fully correlated genetic effects ( $\mathbf{1}_R \mathbf{1}_R^T$ ) and independent genetic effects ( $\mathbf{I}_R$ ). The relation to StructLMM becomes apparent from Eqs (19-20), considering that  $\mathbf{x} \odot \boldsymbol{\beta} = \text{diag}(\mathbf{x})\boldsymbol{\beta}$ . In essence, SKAT-O interpolates between models that consider different correlation structures of genetic effects across variants, whereas StructLMM interpolates between different covariance models for per-individual genetic effects.

A second related model is an LMM-based method for gene-gene ( $G \times G$ ) interaction test introduced in [32], between a single variant and multiple others. While conceptually related, the proposed tests scales quadratically with the number of samples (StructLMM is linear), and hence this method cannot be applied to larger datasets.

Additionally, the method does not consider a joint association test but instead is limited to testing for interactions at this point.

Finally, we note that the approaches we employ for efficient fitting of the null and alternative models in StructLMM borrows ideas from previous efficient implementations of linear mixed models using low-rank random effect, including [5, 33, 7, 12, 30].

### 3. Comparison methods

We compare StructLMM to alternative implementations of single and multi environments G×E models. Additionally, for association tests, we present comparisons to alternative implementations of linear persistent effect tests. Here we provide full details on the models these tests build on. For a compact tabular summary, we refer to **Supp. Table 2** with the methods (SingleEnv-Renv, LMM-Renv and MultiEnv-Renv-LRT) that are used in the main text results displayed in bold. These methods were chosen as the default comparison partners as they use the same random effect term to model the additive environment under the null as StructLMM and hence are most directly comparable.

#### 3.1. Single environment models

**Single Environment model with Single environment additive effect (SingleEnv-Senv).** Standard interaction tests and joint association tests consider the following model [16, 17] for each environment  $e_l$  in isolation:

$$\mathbf{y} \sim \mathcal{N} \left( \mathbf{X}\mathbf{b} + \underbrace{\mathbf{x}\beta_G}_{\text{G}} + \underbrace{(\mathbf{x} \odot \mathbf{e}_l)\beta_{\text{G} \times \text{E}}}_{\text{G} \times \text{E}} + \underbrace{(\mathbf{e}_l)\gamma}_{\text{E}}, \underbrace{\sigma_n^2 \mathbf{I}}_{\text{noise}} \right), \quad (56)$$

where  $\mathbf{y}$  is an  $N \times 1$  phenotype vector for  $N$  individuals,  $\mathbf{X}$  is the  $N \times K$  fixed effect design matrix for  $K$  covariates,  $\mathbf{b}$  is the  $K \times 1$  vector of their effect sizes,  $\mathbf{x}$  is an  $N \times 1$  genotype vector for the variant being tested,  $\beta_G$  the corresponding genetic effect and  $\psi$  the residuals. This model can be used to assess the presence of G×E interaction with environment  $e_l$  by testing  $\beta_{\text{G} \times \text{E}} \neq 0$  (1 df). A joint P value that corresponds to the alternative hypothesis that at least one of  $L$  environments is participating in G×E effects, can then be constructed by performing  $L$  tests followed by appropriate adjustment for multiple testing (we consider Bonferroni). Similarly, a joint association test that accounts for G×E effects due to single environments  $e_l$  can be derived by testing  $[\beta_G, \beta_{\text{G} \times \text{E}}] \neq \mathbf{0}$  (2 df), where again multiple environments and their corresponding tests can be combined using Bonferroni adjustment.

**Single Environment model with multi-environment additive effect as Fixed effect (SingleEnv-Fenv).** One can extend the standard approach presented above by mod-

elling additive environment effects from multiple environments:

$$\mathbf{y} \sim \mathcal{N} \left( \mathbf{X}\mathbf{b} + \underbrace{\mathbf{x}\beta_G}_G + \underbrace{(\mathbf{x} \odot \mathbf{e}_l)\beta_{G \times E}}_{G \times E} + \underbrace{\sum_{l=1}^L (\mathbf{e}_l)\gamma_l}_E, \underbrace{\sigma_n^2 \mathbf{I}}_{\text{noise}} \right). \quad (57)$$

Again, the presence of interaction and association can be assessed by testing  $\beta_{G \times E} \neq 0$  and  $[\beta_G, \beta_{G \times E}] \neq \mathbf{0}$  for each environment respectively, where again multiple environments can be combined using Bonferroni adjustment. This model provides a more accurate null model for both tests, modelling additive effects of other relevant environments.

**Single Environment model with multi-environment additive effect as Random effect (SingleEnv-Renv).** Alternatively, one can use a random effect to model multivariate additive environments, which is analogous to the approach taken in StructLMM. Specifically, one can define an environmental covariance  $\Sigma$  based on the observed environments as described in Section 1.2 and consider the model

$$\mathbf{y} \sim \mathcal{N} \left( \mathbf{X}\mathbf{b} + \underbrace{\mathbf{x}\beta_G}_G + \underbrace{(\mathbf{x} \odot \mathbf{e}_l)\beta_l}_{G \times E}, \underbrace{\sigma_e^2 \Sigma}_E + \underbrace{\sigma_n^2 \mathbf{I}}_{\text{noise}} \right), \quad (58)$$

where again interaction and association tests can be implemented as described above. In the main paper, we consider this model as it has the same background model as StructLMM but perform comparisons with alternative implementations in **Supp. Fig. 2**.

### 3.2. Association tests

**Linear model (LM).** For comparison, we consider a conventional linear model (LM), assuming genetic effect sizes that are constant in the population

$$\mathbf{y} \sim \mathcal{N} \left( \mathbf{X}\mathbf{b} + \underbrace{\mathbf{x}\beta_G}_G, \underbrace{\sigma_n^2 \mathbf{I}}_{\text{noise}} \right). \quad (59)$$

Association with genetic variant  $\mathbf{x}$  can be assessed through the 1dof test  $\beta_G \neq 0$ .

**Linear model with multivariate additive environment effects.** In a linear model, multivariate additive effects from environment can be accounted for by the means of a fixed effect term, which results in the model (LM-Fenv):

$$\mathbf{y} \sim \mathcal{N} \left( \mathbf{X}\mathbf{b} + \underbrace{\mathbf{x}\beta_G}_G + \underbrace{\sum_{l=1}^L (\mathbf{e}_l)\gamma_l}_E, \underbrace{\sigma_n^2 \mathbf{I}}_{\text{noise}} \right). \quad (60)$$

Alternatively, similarly to StructLMM, one can use a random effect with environment covariance  $\Sigma$ , obtaining the following linear mixed model (LMM-Renv)

$$\mathbf{y} \sim \mathcal{N} \left( \mathbf{X}\mathbf{b} + \underbrace{\mathbf{x}\beta_G}_G, \underbrace{\sigma_e^2 \Sigma}_E + \underbrace{\sigma_n^2 \mathbf{I}}_{\text{noise}} \right). \quad (61)$$

In the main paper, we consider the random-effect model as it has the same background model as StructLMM but perform comparisons with alternative implementations in **Supp. Fig. 2**.

### 3.2.1. Alternative multivariate $\mathbf{G} \times \mathbf{E}$ tests based on fixed effects

**Multi Environment model with multi-environment additive effect as Fixed effect (MultiEnv-Fenv).** Whilst we implement StructLMM using a random effect term to model  $\mathbf{G} \times \mathbf{E}$ , an alternative implementation using multiple fixed effects can be considered. Explicitly, denoting with  $\mathbf{e}_1, \dots, \mathbf{e}_L$  the  $N \times 1$  vectors for  $L$  environments, a fixed-effect-based multi-environment model can be cast as

$$\mathbf{y} \sim \mathcal{N} \left( \mathbf{X}\mathbf{b} + \underbrace{\mathbf{x}\beta_G}_G + \underbrace{\sum_{l=1}^L (\mathbf{x} \odot \mathbf{e}_l) \beta_l}_{\mathbf{G} \times \mathbf{E}} + \underbrace{\sum_{l=1}^L (\mathbf{e}_l) \gamma_l}_E, \underbrace{\sigma_n^2 \mathbf{I}}_{\text{noise}} \right), \quad (62)$$

where both interactions due to  $L$  environmental variables and their additive effects are modelled as fixed effects. Within this model, the presence of interactions and associations can be assessed by testing  $[\beta_1, \dots, \beta_L] \neq \mathbf{0}$  ( $L$  df test) and  $[\beta, \beta_1, \dots, \beta_L] \neq \mathbf{0}$  ( $L + 1$  df test), respectively.

We consider both an LR-based and a score-based implementations of the tests (in **Supp. Fig. 1b**), which we respectively refer to as MultiEnv-Fenv-LRT and MultiEnv-Fenv-Score. Briefly, for the score test, we employ the Rao's score test statistic [34]

$$\text{RS} = \mathbf{U}_0^T \mathbf{I}_0^{-1} \mathbf{U}_0, \quad (63)$$

where  $\mathbf{U}_0$  and  $\mathbf{I}_0$  are respectively the gradient and the Fisher Information matrix (FIM) with respect to the tested parameters, computed use MLE under the null<sup>†</sup>. Rao's score test statistic has an asymptotic chi-square distribution with the number

<sup>†</sup>Specifically, for the test  $\beta \neq \mathbf{0}$  in the Gaussian model

$$\mathbf{y} \sim \mathcal{N}(\mathbf{F}\alpha + \mathbf{W}\beta, \mathbf{H}), \quad (64)$$

we have

$$U(\beta) = (\mathbf{y} - \mathbf{F}\alpha_0)^T \mathbf{H}_0^{-1} \mathbf{W}, \quad (65)$$

$$I(\beta) = \mathbf{W}^T \mathbf{H}_0^{-1} \mathbf{W}, \quad (66)$$

where  $\alpha_0$  and  $\mathbf{H}_0$  are MLE of  $\alpha$  and  $\mathbf{H}$  under the null model.

of tested parameters as degrees of freedom.

Whilst the LR test is inflated, the score test is deflated when the number of environments is large relative to the sample size and again highlighting the benefits of the StructLMM tests that have constant number of parameters independent of the number of environments.

**Multi Environment model with multi-environment additive effect as Random effect (MultiEnv-Renv).** As with the single environment and linear model tests, the multivariate additive environment effect can be modelled as a random effect with environment covariance  $\Sigma$ , analogous to the StructLMM model, resulting in the following model

$$\mathbf{y} \sim \mathcal{N} \left( \mathbf{X}\mathbf{b} + \underbrace{\mathbf{x}\beta_G}_G + \underbrace{\sum_{l=1}^L (\mathbf{x} \odot \mathbf{e}_l)\beta_l}_{G \times E}, \underbrace{\sigma_e^2 \Sigma}_E + \underbrace{\sigma_n^2 \mathbf{I}}_{\text{noise}} \right), \quad (67)$$

where again interaction and association tests can be performed as described above. Again as described above both LR tests (MultiEnv-Renv-LRT) and score tests (MultiEnv-Renv-Score) can be employed.

**PC-based fixed effect  $G \times E$  test** Motivated by the observation that multi-environment fixed effect models for  $G \times E$  are in general not calibrated, in particular when analysing larger numbers of environments (**Supp. Fig. 1b**), one can in principle consider PCA to reduce the effect number of environmental variables in the test. The model is identical to the approach described in Eq (67), but is based on the leading  $M$  principal components calculated based on the full set of  $L$  environments, where  $M < L$ . We consider this approach when analysing the eQTL data, where the proxy for environment is high-dimensional **Supp. Fig. 20**.

### 3.3. Original analysis of blood eQTL data

We here provide a brief description of the method used in [35] to identify expression quantitative trait loci whose effect depends on cellular context, Briefly, the authors considered the single-environment model in Eq (51) where  $\mathbf{y}$  is the gene expression of the analysed eQTL gene,  $\mathbf{g}$  its cis eQTL and  $\mathbf{e}$  the environment-gene. Then the following procedure was considered:

- Fit model for each eQTL gene as outcome ( $\mathbf{y}$ ) and each genome-wide gene as environment ( $\mathbf{e}$ );
- Compute the test statistic  $\zeta_l = \sum_i z_{il}^2$  for each environment-gene  $l$ , where  $z_{il}$  is the z-score of the interaction  $G \times E$  term from the model with gene eQTL  $i$  as outcome ( $\mathbf{y}$ ) and gene  $l$  as environment ( $\mathbf{e}$ );

- Define the environment-gene with maximal test statistic as proxy gene and regress out its expression from all gene expression levels.

This procedure was repeated ten times, which led to the identification of ten proxy genes. Each proxy gene was then linked to a specific cell type or molecular pathway called module through various analyses (see [35] for more details). Once the proxy genes and the corresponding modules had been identified, Zhernakova *et al.* considered again the single-environment model in Eq (51) to test for interactions with specific modules at individual eQTL (each eQTL gene as outcome and each proxy gene as environment). The significance threshold for this second stage was assessed through permutations and results reported at 5% FDR.

## 4. Simulations

### 4.1. Environment covariates

In order to mimic realistic distributions in the real data analysis, we considered environmental exposures from UK Biobank data. Specifically, we extracted 33 environmental factors, which include 9 ordinal dietary variables ('Oily fish intake', 'Non-oily fish intake', 'Processed meat intake', 'Poultry intake', 'Beef intake', 'Lamb/mutton intake', 'Pork intake', 'Cheese intake' and 'Salt added to food'), three continuous dietary variables ('Cooked vegetable intake', 'Bread intake', 'Tea intake'), three physical activity variables ('Number of days/week walked 10+ minutes', 'Number of days/week of moderate physical activity 10+ minutes', 'Number of days/week of vigorous physical activity 10+ minutes'), 'Alcohol intake frequency', 'Sleep duration', 'Sleep duration residuals squared', 'Townsend deprivation index', 'Smoking status', 'Time spent watching television', 'Usual walking pace', 'Frequency of friend/family visits', 'Time spend outdoors in summer', 'Time spent outdoors in winter', 'Time spent using computer', 'Nap during day', 'Overall health rating', 'Nitrogen dioxide air pollution 2010', 'Nitrogen oxides air pollution 2010', 'Traffic intensity on the nearest major road', 'Average daytime sound level of noise pollution', 'Average evening sound level of noise pollution'<sup>‡</sup>. For the three continuous dietary and five pollutant variables, we removed values exceeding the 99th percentile. For 'Sleep duration', we removed the top and bottom percentiles and for each individual, calculated the squared deviations from the mean sleep duration, creating environmental variable, 'Squared sleep duration res.' (33<sup>rd</sup> environmental variable). For the four variables ('Time spent watching television', 'Time spent using computer', 'Time spend outdoors in summer', 'Time spent outdoors in winter'), less than 0.5 hours of was encoded as 0.5 and we excluded individuals in the upper and lower percentile. These variables were further augmented by element wise interactions with gender and age (and the inclusion of age itself), resulting in a total of  $L = 100$  environmental variables. We randomly assigned environmental profiles from the 70,282 individuals for which all environmental factors were available (based on the

---

<sup>‡</sup>Environments in blue were also considered in the BMI analysis and were preprocessed as described in the corresponding section.

Interim release) to the genotypes. The environmental covariance was built as a linear covariance from standardised values.

## 4.2. Phenotype simulation strategy

Phenotype data were generated as the sum of a persistent genetic effect from a genetic variant ( $\mathbf{g}$ ), additive environment effects from a set of  $E$  environments ( $\mathbf{e}$ ), G×E effects from a subset of  $L_1 \leq L$  environments ( $\mathbf{i}$ ), a contribution from population structure ( $\mathbf{u}$ ) and iid gaussian noise ( $\boldsymbol{\psi}$ )

$$\mathbf{y} = \mathbf{g} + \mathbf{e} + \mathbf{i} + \mathbf{u} + \boldsymbol{\psi}. \quad (68)$$

The individual contributions from each term as in Eq (68) were simulated as follows:

- **Persistent genetic effect** from a randomly selected genetic variant is rescaled to have sample variance  $(1 - \rho) \cdot v_g$ ;
- **Additive environment effects** are generated as  $\mathbf{e} = \mathbf{E}_E \boldsymbol{\beta}_E$ , where  $\mathbf{E}_E$  is the  $N \times L$  design matrix for  $L$  environments in  $N$  individuals and  $\boldsymbol{\beta}_E$  is generated as  $\boldsymbol{\beta}_E \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$ . Subsequently,  $\mathbf{e}$  is rescaled to have sample variance  $v_e$ ;
- **G×E effects** are generated as  $\mathbf{i} = (\mathbf{E}_{G \times E} \boldsymbol{\beta}_{G \times E}) \otimes \mathbf{x}$ , where  $\mathbf{E}_{G \times E}$  is the  $N \times L_1$  design matrix for  $L_1 < L$  environments in  $N$  individuals and  $\boldsymbol{\beta}_{G \times E}$  is generated as  $\boldsymbol{\beta}_{G \times E} \stackrel{\text{iid}}{\sim} \{+1, -1\}$ . Subsequently,  $\mathbf{i}$  is rescaled to have sample variance  $\rho \cdot v_g$ ;
- **Population structure** is simulated as  $\mathbf{u} = \mathbf{F} \boldsymbol{\beta}_{\text{pop}}$ , where  $\mathbf{F}$  is the  $N \times 10$  matrix of the leading 10 principal components from the realised relatedness matrix and  $\boldsymbol{\beta}_{\text{pop}}$  is generated as  $\boldsymbol{\beta}_{\text{pop}} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$ . Subsequently,  $\mathbf{u}$  is rescaled to have sample variance  $v_{\text{pop}}$ ;
- **Noise** is generated as iid Gaussian and rescaled to have sample variance  $v_n = 1 - v_g - v_e - v_{\text{pop}}$ .

In a first set of simulation experiments, we vary the sample size ( $N$ ), the number of environments ( $L$ ), the percentage of them with G×E effects ( $\pi = \frac{L_1}{L}$ ), the variance explained by the total genetic effect (G+G×E effect,  $v_g$ ), the proportion of the genetic variance due to G×E effects ( $\rho$ ) and the variance explained by additive environmental effects ( $v_e$ ) while we fix the variance explained by population structure ( $v_{\text{pop}} = 40\%$ ). The specific parameter settings and the considered ranges of settings are provided in **Supp. Table 1**. Additionally to these extensive simulations, we also simulated scenarios with gene-exposure correlation, skewed and binary environments. Full details on these additional simulations are given in the next sections.

## 4.3. Gene-exposure correlations

We simulated environments that are subject to a genetic effect from a variant in LD with the variant affecting the phenotype. Specifically, we generated new environments



as

$$\text{vec}(\mathbf{E}) \sim \mathcal{N}(\mathbf{0}, \mathbf{C} \otimes \mathbf{R}), \quad (69)$$

where  $\mathbf{C}$  is the covariance between environments and  $\mathbf{R}$  is the covariance between individuals. We set  $\mathbf{C} = \hat{\mathbf{C}}$  and  $\mathbf{R} = v_x \mathbf{x}_e \mathbf{x}_e^T + (1 - v_x) \hat{\mathbf{R}}$ , where  $\hat{\mathbf{C}}$  and  $\hat{\mathbf{R}}$  are column and row covariances estimated from environments in real data respectively,  $\mathbf{x}_e$  the genotype of the variant correlated with the environments and  $v_x$  the variance explained by this variant on environments. After generation, the environments were standardised and then the phenotype simulation procedure described in Section 4.2 was applied.

In our simulations we varied (i) the LD between the variant correlated with the environments and the one associated with the phenotype and (ii) the average heritability of environments ( $v_x$ ) when the phenotype and environments are affected by the same variant. Calibration was assessed in the extreme scenario of same variant and  $v_x = 0.20\%$ .

#### 4.4. Skewed and binary environments

First, we generated environments as

$$\text{vec}(\mathbf{E}) \sim \mathcal{N}(\mathbf{0}, \hat{\mathbf{C}} \otimes \hat{\mathbf{R}}), \quad (70)$$

where  $\hat{\mathbf{C}}$  and  $\hat{\mathbf{R}}$  are environment and individuals covariances estimated from environments in real data. After generation, the environments were transformed to have non-Gaussian distributions and then the phenotype simulation procedure described in Section 4.2 was applied. To generate skewed distributions, we rank-inverse transformed the generated environments to a Gamma distribution with shape  $a$  and scale 1. We varied the shape  $a$  of all simulated environments. Calibration was assessed in the extreme scenario  $a = 0.1$ . For binary environment scenarios, we binarised a fraction  $f_B$  of the generated environments to have a specific sample frequency  $\nu$ . We varied the fraction  $f_B$  of binary environments (fixing  $\nu = 2\%$ ) and the frequency  $\nu$  (fixing  $f_B=1$ ). Calibration was assessed in the extreme scenario  $f_B = 1$  and  $\nu = 0.5\%$ .

## 5. Analysis of BMI in UK Biobank

### 5.1. Hold-out validation of per-individual genetic effect sizes

For out-of-sample assessment of the allelic predictions, we randomly split the 252,188 individuals into training and testing groups (126,094 individuals). We used the training group to fit the model and estimated allelic effects at the 11 interaction loci (5% FDR-adjusted) within sample, similar to the approach taken on the full dataset. We then identified the 5% (6,305) of individuals with the highest and lowest estimated allelic effects. For these strata of individuals, we use their genotype and phenotype information and test for associations using the LMM-Renv from which we identify a subset of strata that have nominally significant P values ( $P < 0.05$ , **Supp. Fig.**

**17a**). Next, we predicted allelic effects in the test set (**Supp. Fig. 17b**) based on all data in the training fraction and the environmental profiles in the test fraction, again identifying the 5% strata of individuals with the highest and lowest estimated allelic effects. We averaged over the predicted allelic effects of test individuals in the 5% strata (only for significant strata in the training set estimations (**Supp. Fig. 17a**)) to produce a mean predicted allelic effect (yellow crosses, **Supp. Fig. 17b**). We then used the genotypes and phenotypes of test individuals in the top and bottom strata to estimate the genetic effect based on their phenotype data (using LMM-Renv). We compared mean predicted allelic effect with the estimated genetic effect (**Supp. Fig. 17c**) and we compared the mean predicted and estimated effects when considering the GxE component only. For the latter comparison, we subtracted the persistent SNP effect (using the estimate from the StructLMM fit on the training fraction) from the predicted mean effect of the strata. For the LMM-Renv estimated effect, this was achieved by the subtracting the variant effect identified from the LMM-Renv using all individuals in the testing set from the variant effect identified from the LMM-Renv using only individuals in the top and bottom 5% strata (**Supp. Fig. 17d**).

## 6. Analysis of cell-context eQTLs in a large blood cohort

### 6.1. Generation of principal components from SAMtools flagstat and Picard tools

The first eight principal components derived from SAMtools flagstat and Picard tools were calculated from the following fields: GC, Bam.genome total, Bam.exon total, Counts.number detected genes, PF BASES, PF ALIGNED BASES, CODING BASES, UTR BASES, INTRONIC BASES, INTERGENIC BASES, PCT CODING BASES, PCT UTR BASES, PCT INTRONIC BASES, PCT INTERGENIC BASES, PCT MRNA BASES, PCT USABLE BASES, MEDIAN CV COVERAGE, MEDIAN 5PRIME BIAS, MEDIAN 3PRIME BIAS, MEDIAN 5PRIME TO 3PRIME BIAS.

### A. Proof for form of Q

First, let us rewrite  $Q_\rho$

$$Q_\rho = \frac{1}{2} \mathbf{y}^T \mathbf{K}_\rho \mathbf{y} \quad (71)$$

$$= \frac{1}{2} (1 - \rho) \mathbf{y}^T \mathbf{P} \text{diag}(\mathbf{x}) \mathbf{1} \mathbf{1}^T \text{diag}(\mathbf{x}) \mathbf{P} \mathbf{y} + \frac{1}{2} \rho \mathbf{y}^T \mathbf{P} \text{diag}(\mathbf{x}) \mathbf{E} \mathbf{E}^T \text{diag}(\mathbf{x}) \mathbf{P} \mathbf{y} \quad (72)$$

$$= \frac{1}{2} c (1 - \rho) \mathbf{v}^T \mathbf{M} \mathbf{v} + \frac{1}{2} \rho \mathbf{v}^T \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{v} \quad (73)$$

where

$$\mathbf{v} = \mathbf{P}^{\frac{T}{2}} \mathbf{y} \quad (74)$$

$$\mathbf{Z} = \mathbf{P}^{\frac{T}{2}} \text{diag}(\mathbf{x}) \quad (75)$$

$$\mathbf{M} = \frac{1}{c} \mathbf{Z} \mathbf{1} \mathbf{1}^T \mathbf{Z}^T \quad (76)$$

$$c = \mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \mathbf{1} \quad (77)$$

Eq (73) can be equivalently written as

$$Q_\rho = \frac{1}{2} \rho \kappa + \frac{1}{2} \tau_\rho \eta_0 \quad (78)$$

where

$$\kappa = \phi + \xi \quad (79)$$

$$\phi = \mathbf{v}^T (\mathbf{I} - \mathbf{M}) \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T (\mathbf{I} - \mathbf{M}) \mathbf{v} \quad (80)$$

$$\xi = 2 \mathbf{v}^T (\mathbf{I} - \mathbf{M}) \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{M} \mathbf{v} \quad (81)$$

$$\tau_\rho = c(1 - \rho) + \frac{\rho}{c} \mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{Z} \mathbf{1} \quad (82)$$

$$\eta_0 = \mathbf{v}^T \mathbf{M} \mathbf{v} \quad (83)$$

Indeed, replacing Eqs (83) in (78)

$$Q_\rho = \frac{1}{2} \rho \mathbf{v}^T (\mathbf{I} - \mathbf{M}) \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T (\mathbf{I} - \mathbf{M}) \mathbf{v} + \quad (84)$$

$$\rho \mathbf{v}^T (\mathbf{I} - \mathbf{M}) \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{M} \mathbf{v} + \quad (85)$$

$$\left( c(1 - \rho) + \frac{\rho}{c} \mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{Z} \mathbf{1} \right) \mathbf{v}^T \mathbf{M} \mathbf{v} \quad (86)$$

$$= \frac{1}{2} \rho \mathbf{v}^T \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{v} - \rho \mathbf{v}^T \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{M} \mathbf{v} + \frac{1}{2} \rho \mathbf{v}^T \mathbf{M} \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{M} \mathbf{v} + \quad (87)$$

$$+ \rho \mathbf{v}^T \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{M} \mathbf{v} - \rho \mathbf{v}^T \mathbf{M} \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{M} \mathbf{v} + \quad (88)$$

$$\frac{1}{2} c(1 - \rho) \mathbf{v}^T \mathbf{M} \mathbf{v} + \frac{1}{2c} \rho \mathbf{1}^T \mathbf{Z}^T \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{Z} \mathbf{1} \mathbf{v}^T \underbrace{\frac{1}{c} \mathbf{Z} \mathbf{1} \mathbf{1}^T \mathbf{Z}^T}_{\mathbf{M}} \mathbf{v} \quad (89)$$

$$= \frac{1}{2} c(1 - \rho) \mathbf{v}^T \mathbf{M} \mathbf{v} + \frac{1}{2} \rho \mathbf{v}^T \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{v} + \quad (90)$$

$$- \frac{1}{2} \rho \mathbf{v}^T \mathbf{M} \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{M} \mathbf{v} + \frac{1}{2} \rho \mathbf{v}^T \underbrace{\frac{1}{c} \mathbf{Z} \mathbf{1} \mathbf{1}^T \mathbf{Z}^T}_{\mathbf{M}} \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \underbrace{\frac{1}{c} \mathbf{Z} \mathbf{1} \mathbf{1}^T \mathbf{Z}^T}_{\mathbf{M}} \mathbf{v} \quad (91)$$

$$= \frac{1}{2} c(1 - \rho) \mathbf{v}^T \mathbf{M} \mathbf{v} + \frac{1}{2} \rho \mathbf{v}^T \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{v} \quad (92)$$

Denoting with  $\mathbf{C}$  a symmetric matrix, with  $\mathbf{A}$  and  $\mathbf{B}$  two general matrices and suppose  $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ , then we have

$$\mathbf{v}^T \mathbf{C} \mathbf{v} \sim \sum_k w_k \chi_1^2, \text{ where } \mathbf{w} = \text{eigh}(\mathbf{C}) \quad (93)$$

and

$$\mathbb{E}(\mathbf{v}^T \mathbf{A} \mathbf{v}) = \text{tr}(\mathbf{A}) \quad (94)$$

$$\text{Cov}(\mathbf{v}^T \mathbf{A} \mathbf{v}, \mathbf{v}^T \mathbf{B} \mathbf{v}) = \text{tr}(\mathbf{A}(\mathbf{B} + \mathbf{B}^T)) \quad (95)$$

where we used  $\mathbb{E}(\mathbf{v}^T \mathbf{A} \mathbf{v} \mathbf{v}^T \mathbf{B} \mathbf{v}) = \text{tr}(\mathbf{A}(\mathbf{B} + \mathbf{B}^T)) + \text{tr}(\mathbf{A}) \text{tr}(\mathbf{B})$  from the matrix cookbook section 8.2.4. Using these properties and that  $\mathbf{M}\mathbf{M} = \mathbf{M}^\S$ , we have

- $\phi \sim \sum_{k=1}^m \lambda_k \chi_1^2$  where  $\boldsymbol{\lambda} = \text{eigh}(\boldsymbol{\Lambda})$
- $\boldsymbol{\Lambda} = \mathbf{E}^T \mathbf{Z}^T (\mathbf{I} - \mathbf{M}) \mathbf{Z} \mathbf{E}$
- $\mathbb{E}(\xi) = 2\text{tr}((\mathbf{I} - \mathbf{M}) \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{M}) = 0$
- $\text{Var}(\xi) = 4\text{tr}((\mathbf{I} - \mathbf{M}) \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{M} \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T)$
- $\text{Cov}(\xi, \eta_0) = 4\text{tr}((\mathbf{I} - \mathbf{M}) \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{M}) = 0$
- $\text{Cov}(\xi, \phi) = 4\text{tr}((\mathbf{I} - \mathbf{M}) \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T \mathbf{M} (\mathbf{I} - \mathbf{M}) \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T (\mathbf{I} - \mathbf{M})) = 0$
- $\text{Cov}(\phi, \eta_0) = 4\text{tr}((\mathbf{I} - \mathbf{M}) \mathbf{Z} \mathbf{E} \mathbf{E}^T \mathbf{Z}^T (\mathbf{I} - \mathbf{M}) \mathbf{M}) = 0$
- $\eta_0 = \mathbf{v}^T \mathbf{M} \mathbf{v} = (\hat{\mathbf{z}}^T \mathbf{v})^T (\hat{\mathbf{z}}^T \mathbf{v}) \sim \chi_1^2$ , as  $\hat{\mathbf{z}} = \frac{\mathbf{Z}\mathbf{1}}{\|\mathbf{Z}\mathbf{1}\|}$  is a direction.

Using Eqs (74-77) and introducing

$$\boldsymbol{\Lambda}_0 = \mathbf{E}^T \text{diag}(\mathbf{x}) \mathbf{P} \text{diag}(\mathbf{x}) \mathbf{E} \quad (96)$$

$$\boldsymbol{\alpha} = \mathbf{E}^T \text{diag}(\mathbf{x}) \mathbf{P} \mathbf{x} \quad (97)$$

$$c = \mathbf{x} \mathbf{P} \mathbf{x}^T \quad (98)$$

we can rewrite  $\boldsymbol{\Lambda}$ ,  $\text{Var}(\xi)$  and  $\tau_\rho$  as

$$\boldsymbol{\Lambda} = \boldsymbol{\Lambda}_0 - \frac{1}{c} \boldsymbol{\alpha} \boldsymbol{\alpha}^T \quad (99)$$

$$\text{Var}(\xi) = \frac{4}{c} \boldsymbol{\alpha}^T \boldsymbol{\Lambda}_0 \boldsymbol{\alpha} - \frac{4}{c^2} (\boldsymbol{\alpha}^T \boldsymbol{\alpha})^2 \quad (100)$$

$$\tau_\rho = c(1 - \rho) + \frac{\rho}{c} \boldsymbol{\alpha}^T \boldsymbol{\alpha} \quad (101)$$

---

<sup>§</sup>from which follows  $(\mathbf{I} - \mathbf{M})(\mathbf{I} - \mathbf{M}) = (\mathbf{I} - \mathbf{M})$  and  $\mathbf{M}(\mathbf{I} - \mathbf{M}) = 0$

## References

- [1] Rasmussen, C. E. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, 63–71 (Springer, 2004).
- [2] Searle, S. R. & Khuri, A. I. *Matrix algebra useful for statistics* (John Wiley & Sons, 2017).
- [3] Hayes, B. J., Visscher, P. M. & Goddard, M. E. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet Res (Camb)* **91**, 47–60 (2009).
- [4] Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**, 833–835 (2011).
- [5] Lippert, C. *et al.* Fast linear mixed models for genome-wide association studies. *Nat Methods* **8**, 833–5 (2011).
- [6] Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**, 82–93 (2011).
- [7] Lippert, C. *et al.* Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. *Bioinformatics* **30**, 3206–14 (2014).
- [8] Davies, R. B. The distribution of a linear combination of  $\chi^2$  random variables. *Applied Statistics* **29**, 323–333 (1980).
- [9] Liu, H., Tang, Y. & Zhang, H. H. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis* **53**, 853–856 (2009).
- [10] Wu, B., Guan, W. & Pankow, J. S. On Efficient and Accurate Calculation of Significance P-Values for Sequence Kernel Association Testing of Variant Set. *Ann. Hum. Genet.* **80**, 123–35 (2016).
- [11] Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* **91**, 224–37 (2012).
- [12] Casale, F. P., Rakitsch, B., Lippert, C. & Stegle, O. Efficient set tests for the genetic analysis of correlated traits. *Nat Methods* **12**, 755–8 (2015).
- [13] Kostem, E. & Eskin, E. Improving the accuracy and efficiency of partitioning heritability into the contributions of genomic regions. *The American Journal of Human Genetics* **92**, 558–564 (2013).
- [14] Kass, R. E. & Raftery, A. E. Bayes Factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).

- [15] Henderson, C. *Applications of Linear Models in Animal Breeding* (University of Guelph, 1984).
- [16] Gauderman, W. J. *et al.* Update on the State of the Science for Analytical Methods for Gene-Environment Interactions. *Am. J. Epidemiol.* **186**, 762–770 (2017).
- [17] Kraft, P., Yen, Y. C., Stram, D. O., Morrison, J. & Gauderman, W. J. Exploiting gene-environment interaction to detect genetic associations. *Hum. Hered.* **63**, 111–9 (2007).
- [18] Dai, J. Y., Kooperberg, C., Leblanc, M. & Prentice, R. L. Two-stage testing procedures with independent filtering for genome-wide gene-environment interaction. *Biometrika* **99**, 929–944 (2012).
- [19] Chatterjee, N., Kalaylioglu, Z., Moslehi, R., Peters, U. & Wacholder, S. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *Am J Hum Genet* **79**, 1002–16 (2006).
- [20] Jiao, S. *et al.* Sberia: set-based gene-environment interaction test for rare and common variants in complex diseases. *Genet Epidemiol* **37**, 452–64 (2013).
- [21] Li, S. *et al.* Physical Activity Attenuates the Genetic Predisposition to Obesity in 20,000 Men and Women from EPIC-Norfolk Prospective Population Study. *PLoS Med.* **7**, e1000332 (2010).
- [22] Tyrrell, J. *et al.* Gene-obesogenic environment interactions in the UK Biobank study. *Int. J. Epidemiol.* **46**, 559–575 (2017).
- [23] Rask-Andersen, M., Karlsson, T., Ek, W. E. & Johansson, Å. Gene-environment interaction study for BMI reveals interactions between genetic factors and physical activity, alcohol consumption and socioeconomic status. *PLOS Genet.* **13**, e1006977 (2017).
- [24] Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
- [25] Lin, X., Lee, S., Christiani, D. C. & Lin, X. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* **14**, 667–81 (2013).
- [26] Lin, X. *et al.* Test for rare variants by environment interactions in sequencing association studies. *Biometrics* **72**, 156–64 (2016).
- [27] Tzeng, J.-Y. *et al.* Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *Am J Hum Genet* **89**, 277–88 (2011).

- [28] Zhao, G., Marceau, R., Zhang, D. & Tzeng, J.-Y. Assessing gene-environment interactions for common and rare variants with binary traits using gene-trait similarity regression. *Genetics* **199**, 695–710 (2015).
- [29] Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
- [30] Casale, F. P., Horta, D., Rakitsch, B. & Stegle, O. Joint genetic analysis using variant sets reveals polygenic gene-context interactions. *PLoS genetics* **13**, e1006693 (2017).
- [31] Young, A. I., Wauthier, F. & Donnelly, P. Multiple novel gene-by-environment interactions modify the effect of fto variants on body mass index. *Nat Commun* **7**, 12724 (2016).
- [32] Crawford, L., Zeng, P., Mukherjee, S. & Zhou, X. Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLOS Genet.* **13**, e1006869 (2017).
- [33] Listgarten, J. *et al.* A powerful and efficient set test for genetic markers that handles confounding. *Bioinformatics* (2013).
- [34] Rao, C. R. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 44, 50–57 (Cambridge University Press, 1948).
- [35] Zhernakova, D. V. *et al.* Identification of context-dependent expression quantitative trait loci in whole blood. *Nature genetics* **49**, 139–145 (2017).





# Appendix B: Summary of results from StructLMM interaction test described in Chapter 4 for the different considered phenotypes

**Table B1 Loci containing significant interaction effects when examining ten phenotypes** | Table of the 74 loci (the locus numbers referred to in Chapter 4 are displayed in column 1) that contain significant interaction effects (cFDR adjusted  $P < 0.05$ , as described in Section 4.3.4), summarising the interaction and association results for all ten considered phenotypes. Shown are the chromosome (column 2), position (column 3), reference allele (column 4), alternate allele (column 5) and the considered trait (column 6), corresponding to the lead interaction variant (based on the cFDR adjusted  $P$  values) at the locus. The cFDR adjusted  $P$  value (column 7) and the association  $P$  value (column 8) for this lead interaction variant, as well as the lead association  $P$  value for this trait at the considered locus (column 9) with the corresponding position, reference allele and alternate allele (column 10) are also shown. Finally the estimated values of  $\rho$  (which describes the fraction of the genetic variance attributable to  $G \times E$  effects) for significant interaction trait-locus pairs (cFDR adjusted  $P < 0.05$ ) are displayed in column 11. Significant interaction trait-locus pairs (cFDR adjusted  $P < 0.05$ ) are highlighted in yellow, significant association trait-variant pairs ( $P < 5 \times 10^{-8}$  considering the lead interaction variant) are highlighted in green and significant association trait-locus pairs ( $P < 5 \times 10^{-8}$  considering the lead association variant) are highlighted in red.

Locus #	Chrom	Pos	A1	A2	Trait	Lead cFDR P	Assoc P	Lead assoc P	Lead assoc (pos_a1_a2)	p
1	1	78623626	C	T	BMR	0.485115095	9.89E-41	1.50E-53	78450517_C_A	N/A
1	1	78450517	C	A	Body fat %	0.047694988	1.86E-13	1.71E-14	77967507_T_A	0.280916404
1	1	78450517	C	A	Weight	0.122702537	7.04E-49	7.04E-49	78450517_C_A	N/A
1	1	77804242	A	G	BMI	0.314752288	2.4745E-07	6.40E-21	78450517_C_A	N/A
1	1	78623626	C	T	DBP	0.101809306	0.000858182	0.000858182	78623626_C_T	N/A
1	1	79053963	G	A	SBP	0.890856086	0.550761	0.0444988	78515616_T_C	N/A
1	1	77967507	T	A	HC	0.2052569	1.30E-28	2.41E-31	78450517_C_A	N/A
1	1	77804242	A	G	WC	0.235242678	9.10281E-09	2.36E-19	77967507_T_A	N/A
1	1	77989923	T	C	Height	0.512716214	1.88E-03	2.94E-37	78450517_C_A	N/A
1	1	78603463	T	C	PEF	0.80376406	0.922341	5.58218E-05	77958837_A_C	N/A
2	1	177889480	A	G	BMR	0.029553417	3.66E-49	3.66E-49	177889480_A_G	0.092713702
2	1	177889025	A	C	Body fat %	0.001653532	7.41E-33	7.41E-33	177889025_A_C	0.14595585
2	1	177889025	A	C	Weight	7.36785E-05	1.81E-59	9.70E-60	177889480_A_G	0.133170654
2	1	177889480	A	G	BMI	0.000243557	9.15E-58	6.64E-58	177889025_A_C	0.122221004
2	1	177889025	A	C	DBP	0.165634694	0.279346	0.144081	177769090_A_G	N/A
2	1	177902753	G	C	SBP	0.879627989	7.12E-03	0.00556686	177901713_T_C	N/A
2	1	177889025	A	C	HC	1.64901E-05	1.62E-50	1.59E-50	177889480_A_G	0.163157748
2	1	177889025	A	C	WC	0.003604642	1.35E-40	1.35E-40	177889025_A_C	0.130233486
2	1	177904428	T	C	Height	0.977110719	2.68E-05	7.08E-06	177864554_G_A	N/A
2	1	177784137	C	T	PEF	0.926158188	0.936254	1.67E-02	177940341_C_T	N/A
3	1	32129041	A	G	BMR	0.489739316	1.69E-11	1.69E-11	32129041_A_G	N/A
3	1	32129041	A	G	Body fat %	0.13770815	3.65E-03	3.65E-03	32129041_A_G	N/A
3	1	32129041	A	G	Weight	0.033055662	1.22E-09	1.22E-09	32129041_A_G	0.474532888
3	1	32129041	A	G	BMI	0.033303101	4.84E-05	4.84E-05	32129041_A_G	0.692052692
3	1	32093167	G	C	DBP	0.663586177	0.0028325	0.000348579	32135022_T_A	N/A
3	1	32068843	G	A	SBP	0.934568754	0.0284506	0.00807843	32132063_C_T	N/A
3	1	32129041	A	G	HC	0.02452063	3.03E-06	1.51E-07	32147673_C_T	0.618713941
3	1	32129041	A	G	WC	0.020793065	5.80E-06	5.66602E-06	32131163_C_G	0.668326792
3	1	32135022	T	A	Height	0.969035374	7.32211E-08	1.52E-10	32135773_A_G	N/A
3	1	32134763	C	T	PEF	0.869709873	0.000121185	1.14548E-06	32093167_G_C	N/A
4	1	209517898	G	A	BMR	0.61189081	9.91E-09	3.85E-09	209547268_T_C	N/A
4	1	209541991	T	A	Body fat %	0.048494308	2.60E-06	1.24E-06	209520450_T_C	0.626154327
4	1	209543560	T	G	Weight	0.167028228	1.80E-08	8.99E-10	209517989_A_G	N/A
4	1	209519772	G	C	BMI	0.344742013	4.95E-09	9.15825E-10	209520450_T_C	N/A
4	1	209517898	G	A	DBP	0.326251649	0.00400114	0.000571677	209547268_T_C	N/A
4	1	209541832	A	T	SBP	0.938691084	0.0269913	0.00652523	209435733_T_C	N/A
4	1	209519772	G	C	HC	0.212481146	1.42E-09	1.42E-09	209519772_G_C	N/A
4	1	209543560	T	G	WC	0.321064547	1.53E-09	1.54279E-10	209439141_T_C	N/A
4	1	209437172	T	A	Height	0.601509791	5.82E-01	2.89E-01	209547268_T_C	N/A
4	1	209517898	G	A	PEF	0.960817432	0.635425	0.321481	209547268_T_C	N/A
5	1	177841895	G	A	BMR	0.667325375	4.78E-09	4.78E-09	177841895_G_A	N/A
5	1	177841895	G	A	Body fat %	0.451031978	4.80E-04	4.80E-04	177841895_G_A	N/A
5	1	177841895	G	A	Weight	0.236817639	2.22E-09	2.22E-09	177841895_G_A	N/A
5	1	177841895	G	A	BMI	0.407072933	4.02E-06	4.02E-06	177841895_G_A	N/A
5	1	177841895	G	A	DBP	0.828133625	0.0629333	0.0629333	177841895_G_A	N/A
5	1	177841895	G	A	SBP	1	0.911727	0.911727	177841895_G_A	N/A
5	1	177841895	G	A	HC	0.041989446	1.73E-07	1.73E-07	177841895_G_A	0.556811929
5	1	177841895	G	A	WC	0.113713105	1.11E-04	1.11E-04	177841895_G_A	N/A
5	1	177841895	G	A	Height	0.947969629	1.35E-04	1.35E-04	177841895_G_A	N/A
5	1	177841895	G	A	PEF	0.961987137	0.713037	0.713037	177841895_G_A	N/A
6	2	636017	C	T	BMR	0.206172621	2.80E-60	6.86E-61	615627_C_T	N/A
6	2	628749	C	T	Body fat %	0.049673689	2.27E-25	2.27E-25	628749_C_T	0.088949109
6	2	636017	C	T	Weight	0.036935355	2.64E-62	2.16E-62	636790_T_C	0.048659565
6	2	637597	C	T	BMI	0.067251227	9.33026E-58	5.76808E-58	628504_A_G	N/A
6	2	615658	A	G	DBP	0.398428505	3.56E-03	5.32206E-08	651407_T_C	N/A
6	2	603883	G	T	SBP	0.833461782	2.79012E-05	3.35937E-08	651407_T_C	N/A
6	2	651430	T	C	HC	0.348093557	1.07E-49	1.07E-49	651430_T_C	N/A
6	2	651430	T	C	WC	0.005475278	7.32413E-42	7.32413E-42	651430_T_C	0.098096406
6	2	630034	A	T	Height	0.976959777	2.90E-04	7.29E-05	600575_G_T	N/A
6	2	621954	T	C	PEF	0.865597041	0.00571869	8.46E-04	620297_A_G	N/A
7	2	422144	T	C	BMR	0.285740488	3.24E-55	1.64E-56	417167_T_C	N/A
7	2	417167	T	C	Body fat %	0.077745921	3.10E-30	3.10E-30	417167_T_C	N/A
7	2	417167	T	C	Weight	0.012488514	2.47E-62	2.47E-62	417167_T_C	0.056226875
7	2	554731	T	C	BMI	0.035924077	4.86E-09	3.93E-59	417167_T_C	0.326819518

Locus #	Chrom	Pos	A1	A2	Trait	Lead cFDR P	Assoc P	Lead assoc P	Lead assoc (pos_a1_a2)	p
7	2	409112	A	G	DBP	0.704699245	0.181109	0.000153779	430975_C_T	N/A
7	2	409112	A	G	SBP	0.886962052	0.353992	0.00730148	450234_G_A	N/A
7	2	417167	T	C	HC	0.030189074	3.02E-47	3.02E-47	417167_T_C	0.06680839
7	2	417167	T	C	WC	0.039946183	5.43E-42	5.43E-42	417167_T_C	0.064012608
7	2	554731	T	C	Height	0.869953554	3.46E-02	2.99E-04	466003_G_A	N/A
7	2	407713	A	C	PEF	0.83342247	0.00689385	0.00689385	407713_A_C	N/A
8	2	181436641	C	T	BMR	0.66692424	1.20E-08	9.88E-10	181570507_A_G	N/A
8	2	181436641	C	T	Body fat %	0.173257094	3.97E-10	9.35E-11	181568007_A_G	N/A
8	2	181436641	C	T	Weight	0.158451389	4.31E-12	2.48E-13	181570507_A_G	N/A
8	2	181436641	C	T	BMI	0.019256633	6.19699E-12	8.71E-13	181570507_A_G	0.437541334
8	2	181436641	C	T	DBP	0.465213155	0.0309296	0.0309296	181436641_C_T	N/A
8	2	181607676	A	C	SBP	0.957429638	0.026061	0.0096148	181517642_T_C	N/A
8	2	181436641	C	T	HC	0.013059068	1.86E-10	3.82E-12	181517642_T_C	0.515004806
8	2	181561742	T	C	WC	0.085039056	1.02E-13	2.09E-14	181567556_T_A	N/A
8	2	181561742	T	C	Height	0.984961919	2.43E-01	1.47E-02	181605907_A_G	N/A
8	2	181599070	C	A	PEF	0.96306362	0.916364	0.159004	181528663_T_C	N/A
9	3	49936910	T	A	BMR	0.340018606	5.79E-23	1.80E-23	49920571_T_C	N/A
9	3	50041313	C	T	Body fat %	0.165763781	1.37446E-20	1.66E-22	49920571_T_C	N/A
9	3	49938227	C	G	Weight	0.09232712	5.7149E-31	4.00E-32	49920571_T_C	N/A
9	3	49941436	G	A	BMI	0.025648863	1.5507E-36	9.35E-40	49920571_T_C	0.222685097
9	3	49860854	T	C	DBP	0.048316304	1.41405E-06	2.38187E-11	49881134_A_T	0.512688295
9	3	49860854	T	C	SBP	0.820360422	0.000008811	0.000008811	49860854_T_C	N/A
9	3	49941436	G	A	HC	0.068382623	4.08753E-15	1.60E-17	49920571_T_C	N/A
9	3	50210289	C	G	WC	0.129612367	8.68969E-16	2.68E-22	49920571_T_C	N/A
9	3	50198537	A	G	Height	0.8622064	8.52E-01	0.000227307	50071965_C_T	N/A
9	3	49843723	T	C	PEF	0.832789753	0.00042064	0.00042064	49843723_T_C	N/A
10	4	106263450	T	G	BMR	0.108323236	1.23E-19	9.18E-24	106211443_A_G	N/A
10	4	106230708	T	C	Body fat %	0.41634082	0.182115	6.58E-04	106181573_G_A	N/A
10	4	106293180	G	A	Weight	0.041020046	4.29134E-06	1.02E-10	106216667_A_G	0.616611037
10	4	106293180	G	A	BMI	0.274934427	0.99075	0.00155541	106044276_C_A	N/A
10	4	106181573	G	A	DBP	0.123488293	0.055048	0.0539094	106189614_G_T	N/A
10	4	106263450	T	G	SBP	0.795080238	2.40696E-06	4.77067E-07	106259572_T_C	N/A
10	4	106293180	G	A	HC	0.060676364	0.00129835	4.40E-06	106044276_C_A	N/A
10	4	106044283	T	A	WC	0.327194132	1.45E-02	0.00027627	106081636_T_C	N/A
10	4	106181570	G	A	Height	0.680839511	7.88672E-23	1.86E-50	106216367_T_C	N/A
10	4	106257007	A	C	PEF	0.771439027	9.22E-04	6.68413E-07	106082120_G_A	N/A
11	4	102708997	C	A	BMR	0.488783312	2.42E-09	2.42E-09	102708997_C_A	N/A
11	4	102708997	C	A	Body fat %	0.159552043	1.07E-11	1.07E-11	102708997_C_A	N/A
11	4	102708997	C	A	Weight	0.093691207	6.98E-13	6.98E-13	102708997_C_A	N/A
11	4	102708997	C	A	BMI	0.019585217	1.40E-19	1.40E-19	102708997_C_A	0.247258419
11	4	102709308	T	C	DBP	0.904262715	0.0266914	0.0266914	102709308_T_C	N/A
11	4	102708997	C	A	SBP	0.950617421	0.202831	0.0397945	102709308_T_C	N/A
11	4	102708997	C	A	HC	0.504370621	3.50E-12	3.50E-12	102708997_C_A	N/A
11	4	102708997	C	A	WC	0.15748701	4.15E-11	4.15E-11	102708997_C_A	N/A
11	4	102708997	C	A	Height	0.986225006	1.37E-03	1.37E-03	102708997_C_A	N/A
11	4	102709308	T	C	PEF	0.973562014	0.708587	0.708587	102709308_T_C	N/A
12	5	87497612	T	C	BMR	0.515889613	4.95E-09	6.75E-18	87682877_C_G	N/A
12	5	87497612	T	C	Body fat %	0.194829457	3.19726E-05	1.32E-10	87682877_C_G	N/A
12	5	87497612	T	C	Weight	0.043407349	3.07E-09	8.10E-20	87682877_C_G	0.621358715
12	5	87840688	G	A	BMI	0.082757387	7.67395E-10	7.96E-18	87682877_C_G	N/A
12	5	88248730	A	C	DBP	0.075404794	0.00767108	5.0774E-06	87301956_G_C	N/A
12	5	87512532	G	A	SBP	0.857420733	0.000984105	5.59374E-05	87393746_A_G	N/A
12	5	87497612	T	C	HC	0.149863675	8.52E-06	4.56E-12	87682877_C_G	N/A
12	5	87497612	T	C	WC	0.046965639	1.391E-07	5.55E-17	87682877_C_G	0.619896312
12	5	87997279	A	G	Height	0.982744622	8.06E-01	1.52E-03	87469449_C_A	N/A
12	5	87795525	G	A	PEF	0.88067218	0.00408832	1.50E-03	87328066_C_T	N/A
13	5	6712834	T	C	BMR	0.847309613	4.53E-08	8.75E-09	6754402_T_C	N/A
13	5	6739039	T	G	Body fat %	0.011994174	1.16E-04	5.28E-05	6754402_T_C	0.846513595
13	5	6739039	T	G	Weight	0.255630492	3.50E-09	1.04E-09	6754402_T_C	N/A
13	5	6737571	A	G	BMI	0.126467521	0.00149388	6.33E-04	6754402_T_C	N/A
13	5	6711630	C	T	DBP	0.794477911	0.0652209	0.0378624	6754402_T_C	N/A
13	5	6712834	T	C	SBP	0.991428861	0.572664	0.358811	6745571_A_T	N/A
13	5	6739039	T	G	HC	0.437364134	5.54E-06	1.77E-06	6745571_A_T	N/A
13	5	6712834	T	C	WC	0.210424144	5.17E-05	1.23E-05	6754402_T_C	N/A

Locus #	Chrom	Pos	A1	A2	Trait	Lead cFDR P	Assoc P	Lead assoc P	Lead assoc (pos_a1_a2)	p
13	5	6748659	T	C	Height	0.961025417	6.96E-10	3.27E-11	6712044_C_G	N/A
13	5	6718563	A	G	PEF	0.994713978	0.150894	0.130009	6749085_T_G	N/A
14	6	34717578	C	T	BMR	0.045232701	2.43E-68	1.58E-72	34589632_C_T	0.099028865
14	6	35238183	T	G	Body fat %	0.077080633	3.94127E-08	1.7804E-25	34735883_T_C	N/A
14	6	35238183	T	G	Weight	0.063112733	1.45E-21	4.58E-68	34552736_G_A	N/A
14	6	34769765	C	T	BMI	0.177891935	6.71429E-24	1.58689E-29	34644261_G_A	N/A
14	6	34545999	A	C	DBP	0.400660371	0.0674444	9.90266E-05	34796084_C_T	N/A
14	6	34478514	C	T	SBP	0.872543695	0.48834	0.000123508	35044456_A_C	N/A
14	6	35163949	T	A	HC	0.159657368	2.41E-19	4.47E-57	34552736_G_A	N/A
14	6	35099094	C	T	WC	0.031996066	5.40519E-16	1.11E-28	34552736_G_A	0.354652517
14	6	34692585	A	G	Height	0.942464965	3.36E-17	1.27E-82	34623905_C_T	N/A
14	6	34506397	T	C	PEF	0.727350582	0.000142915	3.25943E-05	35386872_C_T	N/A
15	6	130378833	T	C	BMR	0.210928502	8.14E-53	8.24E-56	130374461_T_A	N/A
15	6	130350294	G	A	Body fat %	0.313762326	8.53E-02	1.28E-02	130440798_G_A	N/A
15	6	130350294	G	A	Weight	0.095798826	8.98E-33	3.62E-33	130384187_C_T	N/A
15	6	130350294	G	A	BMI	0.370933604	8.31392E-06	4.97E-06	130384187_C_T	N/A
15	6	130445031	G	A	DBP	0.490453734	0.924806	0.0312236	130335109_C_T	N/A
15	6	130335109	C	T	SBP	0.855806143	9.27E-04	9.27E-04	130335109_C_T	N/A
15	6	130350294	G	A	HC	0.034810144	6.28E-18	7.90E-19	130384187_C_T	0.258647546
15	6	130349119	C	T	WC	0.319902097	6.24E-04	1.83E-04	130384187_C_T	N/A
15	6	130377843	C	T	Height	0.949991683	3.71E-70	9.57E-78	130374461_T_A	N/A
15	6	130420941	C	T	PEF	0.79027379	0.000597738	1.03087E-05	130345835_G_A	N/A
16	6	31597700	C	T	BMR	0.023965939	1.84E-38	4.58E-47	31840021_T_A	0.203315122
16	6	31244789	A	T	Body fat %	0.06490973	0.340236	1.35456E-10	32071637_T_C	N/A
16	6	31323416	G	C	Weight	0.000774273	6.22E-18	3.59E-32	31613739_T_C	0.387742637
16	6	31346621	G	C	BMI	0.014350677	8.36205E-05	1.41953E-17	31613739_T_C	0.79177428
16	6	31319478	A	G	DBP	0.011875155	7.52007E-14	5.52164E-19	31242089_A_T	0.360641786
16	6	31178465	A	C	SBP	0.109435311	0.000129852	2.8755E-15	31610686_A_G	N/A
16	6	31316526	T	C	HC	0.001723726	1.84E-13	3.45603E-27	31613739_T_C	0.389986122
16	6	31236854	T	C	WC	0.005669761	0.56865	5.59033E-11	31613739_T_C	0.971235159
16	6	32623223	A	C	Height	0.306589401	1.38448E-29	2.22E-80	31327895_G_A	N/A
16	6	32107851	G	A	PEF	0.723450255	0.000576952	3.28001E-11	31147476_G_A	N/A
17	6	126964675	G	A	BMR	0.110429772	1.37E-31	1.17E-38	126704795_C_T	N/A
17	6	126760994	T	C	Body fat %	0.06057023	1.9727E-11	3.32E-12	127048230_C_T	N/A
17	6	126964675	G	A	Weight	0.038994932	1.73E-07	4.49E-11	126704795_C_T	0.383873185
17	6	126972883	A	G	BMI	0.091025464	0.000192213	9.01251E-07	127048609_T_A	N/A
17	6	126717064	A	G	DBP	0.125990236	0.00511951	4.10E-10	126938446_C_T	N/A
17	6	127183470	A	G	SBP	0.800604023	3.35084E-06	4.98322E-09	126938446_C_T	N/A
17	6	126975115	A	G	HC	0.37336374	0.0284294	1.62E-07	126604298_G_A	N/A
17	6	127118646	G	A	WC	0.49437224	0.0217578	5.58E-04	127048230_C_T	N/A
17	6	127149538	C	T	Height	0.363609625	1.40E-17	2.30E-98	126851160_C_T	N/A
17	6	126802598	A	G	PEF	0.858070471	1.75E-04	8.61235E-05	126717064_A_G	N/A
18	6	108983527	C	T	BMR	0.190879657	6.77E-31	1.69E-33	108876002_G_A	N/A
18	6	108998953	C	T	Body fat %	0.32412585	5.54646E-06	7.9093E-10	108865663_T_G	N/A
18	6	108983527	C	T	Weight	0.123876819	1.80E-23	4.28E-28	108876002_G_A	N/A
18	6	109020634	T	C	BMI	0.023087066	0.00450541	3.399E-12	108888593_C_G	0.922163258
18	6	109009194	A	G	DBP	0.837929337	0.40067	0.00686157	109009646_C_T	N/A
18	6	108994161	A	G	SBP	0.844286939	0.000782267	2.19841E-05	109009646_C_T	N/A
18	6	108865663	T	G	HC	0.105530283	2.30E-13	1.26E-16	108888593_C_G	N/A
18	6	109005588	A	G	WC	0.064037441	4.65E-12	8.86E-18	108888593_C_G	N/A
18	6	109020634	T	C	Height	0.885430683	2.99E-10	6.54E-24	108983527_C_T	N/A
18	6	109019721	G	A	PEF	0.884334209	6.53028E-05	9.47401E-08	108974098_C_T	N/A
19	6	35420923	C	T	BMR	0.531798593	4.07E-24	6.96E-28	35416039_C_A	N/A
19	6	35448189	T	C	Body fat %	0.198488973	6.02E-03	2.69E-03	35424188_A_G	N/A
19	6	35420628	A	C	Weight	0.181329177	6.89E-17	5.19E-19	35448189_T_C	N/A
19	6	35448189	T	C	BMI	0.275373558	1.28302E-06	1.28302E-06	35448189_T_C	N/A
19	6	35120374	C	A	DBP	0.030374146	2.46E-05	3.49E-06	35389031_T_C	0.62344561
19	6	35404354	T	C	SBP	0.800461854	3.70241E-06	1.00941E-07	35388758_C_T	N/A
19	6	35420628	A	C	HC	0.163488335	1.18E-11	4.51E-13	35448189_T_C	N/A
19	6	35420923	C	T	WC	0.194985535	1.10E-07	7.74E-09	35424188_A_G	N/A
19	6	35439932	T	C	Height	0.882397333	3.78E-25	7.76E-59	35395010_C_T	N/A
19	6	35291638	T	C	PEF	0.787027347	1.20E-03	7.07186E-10	35033854_G_A	N/A
20	6	28758828	G	C	BMR	0.366931394	2.33E-19	1.24E-25	29355148_C_A	N/A
20	6	28679945	A	C	Body fat %	0.168892074	7.45E-04	6.45985E-09	29915061_T_C	N/A

Locus #	Chrom	Pos	A1	A2	Trait	Lead cFDR P	Assoc P	Lead assoc P	Lead assoc (pos_a1_a2)	p
20	6	29730185	G	A	Weight	0.088324303	1.20E-09	8.58E-23	29355148_C_A	N/A
20	6	28890800	G	C	BMI	0.044005465	3.21143E-05	1.18735E-06	28736484_C_T	0.760982378
20	6	29354799	A	G	DBP	0.185169245	9.89E-05	1.28483E-07	29730185_G_A	N/A
20	6	29181314	G	A	SBP	0.844210072	0.00111708	0.00111708	29181314_G_A	N/A
20	6	29516242	A	C	HC	0.079355862	2.13631E-08	2.64E-17	29355148_C_A	N/A
20	6	28890800	G	C	WC	0.106376755	5.06217E-05	5.85E-07	29355148_C_A	N/A
20	6	29396874	T	C	Height	0.698566591	6.50E-27	3.07E-40	28908612_A_G	N/A
20	6	28682763	T	C	PEF	0.863122697	0.000856195	3.77856E-10	29730185_G_A	N/A
21	6	31373260	G	C	BMR	0.262411855	1.09E-15	4.66E-22	31213413_T_G	N/A
21	6	31243767	G	C	Body fat %	0.151592573	0.0125029	0.000751218	31164224_C_G	N/A
21	6	31240692	G	A	Weight	0.087015973	2.75E-12	2.33E-14	31213413_T_G	N/A
21	6	31243767	G	C	BMI	0.148286866	0.15347	6.56859E-07	31183907_A_G	N/A
21	6	31411321	A	C	DBP	0.113932783	0.0524721	3.86624E-09	31353639_C_T	N/A
21	6	31170066	A	G	SBP	0.869176808	0.454701	0.00304827	31266085_C_G	N/A
21	6	31373260	G	C	HC	0.076815582	5.74E-06	8.19E-10	31274582_G_A	N/A
21	6	31243767	G	C	WC	0.015729489	0.181094	1.0536E-05	31136575_T_C	0.99198901
21	6	31114335	A	G	Height	0.518795604	4.51E-09	1.03E-62	31331500_C_G	N/A
21	6	31239407	G	T	PEF	0.847194449	0.00303719	2.13951E-08	31022266_G_C	N/A
22	6	32206465	G	A	BMR	0.238218058	1.78E-17	9.93E-22	32451888_A_G	N/A
22	6	32554197	T	G	Body fat %	0.191055512	0.0104197	3.07E-05	32190390_T_G	N/A
22	6	32554197	T	G	Weight	0.066600585	1.03E-08	3.18059E-17	32623371_C_T	N/A
22	6	32554197	T	G	BMI	0.111898074	6.18548E-07	2.35877E-08	32644553_C_T	N/A
22	6	32206539	C	T	DBP	0.025690663	7.3314E-07	4.56939E-14	32844103_A_G	0.552141895
22	6	32722754	G	A	SBP	0.885863373	0.352655	1.45418E-07	31800868_G_C	N/A
22	6	31742067	A	G	HC	0.118217725	1.06895E-05	7.80542E-13	32644553_C_T	N/A
22	6	31722780	C	T	WC	0.159642205	0.410621	1.93502E-06	32644553_C_T	N/A
22	6	32634318	C	A	Height	0.437590146	9.72E-26	2.66E-33	32652620_G_A	N/A
22	6	32754091	C	T	PEF	0.814638209	0.000414469	6.87804E-09	32627934_G_A	N/A
23	6	31101674	C	T	BMR	0.506229064	1.71E-14	9.65E-19	31117075_C_A	N/A
23	6	31098134	A	T	Body fat %	0.171479518	4.88E-01	0.000280281	31248886_C_T	N/A
23	6	31101674	C	T	Weight	0.064955614	1.48E-10	5.41E-13	31113214_C_A	N/A
23	6	31243008	G	T	BMI	0.154508909	0.125087	2.53456E-05	31248886_C_T	N/A
23	6	31242762	C	T	DBP	0.149548279	3.67297E-05	4.55333E-08	31117075_C_A	N/A
23	6	31136666	A	G	SBP	0.817042645	0.000916325	1.07217E-05	31141836_C_T	N/A
23	6	31101674	C	T	HC	0.094927732	5.56E-07	4.41E-12	31113214_C_A	N/A
23	6	31101674	C	T	WC	0.016024078	1.02E-01	4.87E-04	31113214_C_A	0.96679765
23	6	31286247	G	A	Height	0.964707584	3.46E-09	3.41E-27	31117075_C_A	N/A
23	6	31130502	T	C	PEF	0.87751256	0.00712665	6.53E-05	31299899_T_C	N/A
24	6	31382717	A	T	BMR	0.092644371	9.42E-10	1.32E-18	30914843_C_T	N/A
24	6	29924831	T	C	Body fat %	0.02388548	1.07306E-06	1.07306E-06	29924831_T_C	0.620235603
24	6	30726939	A	C	Weight	0.004871038	2.07243E-12	1.35E-16	30914843_C_T	0.56825856
24	6	29939854	C	T	BMI	0.042059826	6.9938E-07	6.9938E-07	29939854_C_T	0.7523845
24	6	30792117	G	C	DBP	0.012497354	2.02411E-09	2.36904E-12	30342753_G_A	0.478461664
24	6	30888161	T	C	SBP	0.853944211	0.000573371	0.000264558	31363026_G_T	N/A
24	6	31330015	C	G	HC	0.004140969	2.73E-07	1.14968E-11	29939854_C_T	0.560855903
24	6	30726939	A	C	WC	0.002287469	0.042813	0.00476112	31329386_A_G	0.990512423
24	6	31382717	A	T	Height	0.53011149	5.4219E-22	4.77E-26	30811265_G_C	N/A
24	6	31363026	G	T	PEF	0.856748437	0.00265285	2.62904E-12	30285312_G_C	N/A
25	6	40369081	C	T	BMR	0.369344655	1.24E-15	1.24E-15	40369081_C_T	N/A
25	6	40368860	T	G	Body fat %	0.185842807	4.24E-12	9.25E-19	40362023_C_T	N/A
25	6	40369081	C	T	Weight	0.074770749	1.06E-20	1.06E-20	40369081_C_T	N/A
25	6	40369081	C	T	BMI	0.012430743	1.72E-24	7.59E-25	40362751_G_A	0.268451621
25	6	40360755	G	A	DBP	0.215090965	0.0325912	0.00834341	40409243_T_C	N/A
25	6	40362525	C	T	SBP	0.939545451	0.00918073	0.00918073	40362525_C_T	N/A
25	6	40369081	C	T	HC	0.012246303	7.03E-17	1.01E-18	40409243_T_C	0.36645827
25	6	40369081	C	T	WC	0.000443226	6.30E-23	6.30E-23	40369081_C_T	0.341530569
25	6	40369159	A	G	Height	0.987785052	4.22E-01	4.74E-02	40362751_G_A	N/A
25	6	40409243	T	C	PEF	0.93261618	0.573655	0.186134	40362751_G_A	N/A
26	6	31380449	C	T	BMR	0.517486079	3.88E-13	9.62E-15	31046494_G_A	N/A
26	6	31242174	G	A	Body fat %	0.25316997	3.01E-01	8.89E-04	31242509_C_T	N/A
26	6	31242174	G	A	Weight	0.051791262	3.23E-06	1.16E-09	31316079_A_C	N/A
26	6	31242174	G	A	BMI	0.164608554	0.520013	3.70E-05	31327064_T_C	N/A
26	6	31242174	G	A	DBP	0.132763883	9.88915E-06	4.21451E-07	31379817_T_C	N/A
26	6	31270084	A	G	SBP	0.848005549	4.65632E-06	1.31848E-06	31270176_T_C	N/A



Locus #	Chrom	Pos	A1	A2	Trait	Lead cFDR P	Assoc P	Lead assoc P	Lead assoc (pos_a1_a2)	p
26	6	31242174	G	A	HC	0.027600954	1.58E-03	2.53E-10	31242509_C_T	0.748569529
26	6	31242174	G	A	WC	0.025474695	1.82E-01	5.43E-04	31084639_C_T	0.994967675
26	6	31234602	T	C	Height	0.960625411	1.08E-09	8.49E-43	31380449_C_T	N/A
26	6	31234602	T	C	PEF	0.870747881	0.00468511	1.89524E-06	31172332_C_A	N/A
27	6	28453102	G	C	BMR	0.428237182	3.37E-10	6.89E-13	28227145_G_C	N/A
27	6	28514583	T	C	Body fat %	0.287559533	3.46E-03	2.41387E-07	28280137_T_C	N/A
27	6	28453102	G	C	Weight	0.050827235	1.56E-10	1.38E-13	28227145_G_C	N/A
27	6	28465355	C	T	BMI	0.163563889	0.0141103	0.000258903	28103473_G_A	N/A
27	6	28029378	C	G	DBP	0.304910591	1.02375E-07	3.27959E-09	28323968_G_A	N/A
27	6	28495240	T	G	SBP	0.878614784	0.00165133	0.0011838	28447519_T_G	N/A
27	6	28453102	G	C	HC	0.021501267	6.50E-11	6.01E-14	28227145_G_C	0.45715365
27	6	28450437	G	A	WC	0.136838222	0.00412506	0.000442475	28182896_C_G	N/A
27	6	28451144	G	T	Height	0.967576509	4.28E-12	5.53E-21	28465355_C_T	N/A
27	6	28575172	T	A	PEF	0.946066096	4.76E-02	0.00551043	28514583_T_C	N/A
28	6	29849687	G	C	BMR	0.035385433	4.77E-09	2.82E-11	29906313_A_C	0.666752006
28	6	29849657	T	C	Body fat %	0.006765242	2.74E-05	2.74E-05	29849657_T_C	0.723972591
28	6	29849657	T	C	Weight	0.004296876	3.65E-10	2.79E-11	29906313_A_C	0.574739348
28	6	29849657	T	C	BMI	0.027264995	4.19459E-08	4.19459E-08	29849657_T_C	0.615616353
28	6	29849657	T	C	DBP	0.056244164	9.75669E-05	2.62267E-06	30073847_C_T	N/A
28	6	29906978	G	A	SBP	0.915890656	0.00453372	0.00394459	29906313_A_C	N/A
28	6	29842444	G	A	HC	0.001162084	4.86E-05	7.30E-11	29849657_T_C	0.773088523
28	6	29849657	T	C	WC	0.000282912	2.62E-03	1.55E-04	29924159_C_G	0.9515344
28	6	30065319	G	A	Height	0.963089921	4.91E-11	3.78E-21	30073847_C_T	N/A
28	6	30073847	C	T	PEF	0.937658413	0.670551	5.82844E-10	29849657_T_C	N/A
29	6	29546799	G	A	BMR	0.472286604	4.83E-08	2.29E-10	29488249_A_G	N/A
29	6	29546799	G	A	Body fat %	0.166876846	7.00E-04	6.31E-04	29548089_G_A	N/A
29	6	29548089	G	A	Weight	0.014199202	1.35E-09	5.44E-10	29537224_A_G	0.725397337
29	6	29548089	G	A	BMI	0.10675527	1.76E-04	1.76E-04	29548089_G_A	N/A
29	6	29559238	T	C	DBP	0.01403405	4.7084E-10	8.48396E-11	29752808_A_G	0.451037419
29	6	29548089	G	A	SBP	0.865152649	0.505975	0.393688	29752808_A_G	N/A
29	6	29548089	G	A	HC	0.158266856	1.22E-07	9.56E-08	29752808_A_G	N/A
29	6	29548089	G	A	WC	0.0046656	8.93E-02	8.93E-02	29548089_G_A	0.999670769
29	6	29559238	T	C	Height	0.979359506	3.32E-10	1.23E-16	29488249_A_G	N/A
29	6	29488249	A	G	PEF	0.952316331	0.622958	7.14268E-11	29607101_T_C	N/A
30	6	154392675	T	C	BMR	0.374434111	1.15E-09	2.95E-10	154336892_T_C	N/A
30	6	154333183	T	C	Body fat %	0.07685939	2.8525E-07	1.03525E-07	154319449_G_T	N/A
30	6	154392675	T	C	Weight	0.046737778	2.74E-10	7.28E-13	154336892_T_C	0.463803923
30	6	154403818	C	T	BMI	0.044223838	3.23995E-09	1.93E-12	154336892_T_C	0.476470073
30	6	154336892	T	C	DBP	0.683932488	0.802297	0.179087	154333183_T_C	N/A
30	6	154382367	C	T	SBP	0.898095547	0.684383	0.681579	154388306_T_G	N/A
30	6	154333183	T	C	HC	0.028927172	6.03355E-12	1.14E-12	154336892_T_C	0.359303048
30	6	154363387	C	T	WC	0.158115035	2.73173E-08	8.80E-10	154336892_T_C	N/A
30	6	154336892	T	C	Height	0.995767726	2.14E-01	4.90E-02	154405730_A_T	N/A
30	6	154362254	A	G	PEF	0.963148326	0.679429	0.47093	154290321_G_A	N/A
31	6	28578286	A	G	BMR	0.710367858	4.93E-09	3.54E-10	28191057_C_T	N/A
31	6	28578286	A	G	Body fat %	0.104914536	6.51E-13	6.51E-13	28578286_A_G	N/A
31	6	28578286	A	G	Weight	0.091510394	2.87E-15	2.87E-15	28578286_A_G	N/A
31	6	28578286	A	G	BMI	0.18031361	3.33941E-14	3.33941E-14	28578286_A_G	N/A
31	6	28578286	A	G	DBP	0.031492001	6.05656E-06	1.70014E-09	28191057_C_T	0.657198915
31	6	28599105	G	A	SBP	0.892903738	0.683005	0.346436	28228342_A_G	N/A
31	6	28578286	A	G	HC	0.433892653	1.55E-13	1.55E-13	28578286_A_G	N/A
31	6	28578286	A	G	WC	0.016655446	1.61E-06	1.61E-06	28578286_A_G	0.830685889
31	6	28168343	C	T	Height	0.96911031	1.36E-06	4.22E-07	28203300_A_G	N/A
31	6	28207200	C	T	PEF	0.871909334	0.000137277	1.93223E-15	28578286_A_G	N/A
32	6	31440014	C	A	BMR	0.564518084	2.34E-09	6.23E-10	31351572_G_T	N/A
32	6	31440014	C	A	Body fat %	0.072886958	1.96E-02	1.96E-02	31440014_C_A	N/A
32	6	31440014	C	A	Weight	0.042123263	1.30E-08	1.30E-08	31440014_C_A	0.490767893
32	6	31440014	C	A	BMI	0.121497246	0.000112735	0.000112735	31440014_C_A	N/A
32	6	31351572	G	T	DBP	0.394826063	0.000830002	0.00019397	31208591_G_T	N/A
32	6	31351572	G	T	SBP	0.869880608	0.00138061	0.00115922	31220752_C_T	N/A
32	6	31440014	C	A	HC	0.034388632	3.11E-06	3.11E-06	31440014_C_A	0.512832757
32	6	31440014	C	A	WC	0.062811542	4.46E-03	4.46E-03	31440014_C_A	N/A
32	6	31221396	A	G	Height	1	6.18E-24	1.77E-25	31351572_G_T	N/A
32	6	31220752	C	T	PEF	0.973905495	0.854468	0.425266	31351572_G_T	N/A

Locus #	Chrom	Pos	A1	A2	Trait	Lead cFDR P	Assoc P	Lead assoc P	Lead assoc (pos_a1_a2)	p
33	6	29842451	A	G	BMR	0.451164348	8.32E-09	8.32E-09	29842451_A_G	N/A
33	6	29818726	A	G	Body fat %	0.029978553	3.63089E-07	3.63089E-07	29818726_A_G	0.512682676
33	6	29842451	A	G	Weight	0.007591409	1.80E-11	1.80E-11	29842451_A_G	0.597074492
33	6	29842451	A	G	BMI	0.055933008	5.73E-07	5.73E-07	29842451_A_G	N/A
33	6	29907961	G	C	DBP	0.023999641	5.57889E-07	4.19E-07	29818568_T_C	0.461974752
33	6	29834472	C	T	SBP	0.977640435	0.311611	0.0129255	30800068_G_A	N/A
33	6	29821937	C	T	HC	0.016166886	2.98E-11	5.07E-12	29818726_A_G	0.491553622
33	6	29842451	A	G	WC	0.002260426	2.97E-03	2.97E-03	29842451_A_G	0.959382266
33	6	29821937	C	T	Height	0.831377943	2.85E-04	6.26E-16	30800577_C_T	N/A
33	6	30800577	C	T	PEF	1	4.52E-11	6.77262E-12	29818726_A_G	N/A
34	7	92327026	A	G	BMR	0.019834678	6.95E-51	1.65E-55	92250140_T_G	0.117735009
34	7	92277030	T	C	Body fat %	0.216422699	0.0293858	1.02E-02	92253972_G_A	N/A
34	7	92327026	A	G	Weight	0.017180925	3.40E-29	4.36E-33	92253972_G_A	0.208187367
34	7	92327026	A	G	BMI	0.061219812	0.0973123	0.022842	92279363_T_C	N/A
34	7	92223957	T	C	DBP	0.076620722	0.00346198	0.00146737	92237396_A_T	N/A
34	7	92215430	G	A	SBP	0.737655006	1.79784E-09	2.84715E-15	92253972_G_A	N/A
34	7	92327026	A	G	HC	0.095783184	5.30E-19	7.81E-20	92253972_G_A	N/A
34	7	92327026	A	G	WC	0.30562912	4.83E-08	9.63E-09	92253972_G_A	N/A
34	7	92253972	G	A	Height	0.845623833	3.16E-111	5.37E-115	92250140_T_G	N/A
34	7	92243719	C	T	PEF	0.845008254	0.0011117	0.00062571	92237426_G_A	N/A
35	7	2887195	C	T	BMR	0.112324753	7.57E-35	5.22E-40	2859847_G_A	N/A
35	7	2864706	C	A	Body fat %	0.056146988	3.28E-08	1.8973E-09	2912928_T_C	N/A
35	7	2890774	C	T	Weight	0.030382729	3.56803E-30	1.98E-34	2862542_C_A	0.165510044
35	7	2758158	G	A	BMI	0.169158844	0.46547	0.00358741	2912928_T_C	N/A
35	7	2830498	C	T	DBP	0.071489207	0.00459333	0.00402076	2832973_G_A	N/A
35	7	2799686	G	T	SBP	0.888337698	0.323813	0.0101224	2751285_T_G	N/A
35	7	2772431	A	C	HC	0.00392662	2.27E-22	1.14E-26	2862542_C_A	0.186535382
35	7	2772431	A	C	WC	0.021629074	2.46E-09	6.60E-12	2862542_C_A	0.431995909
35	7	2913736	G	T	Height	0.77515197	3.41E-26	1.00E-119	2802522_T_C	N/A
35	7	2811543	T	C	PEF	0.923377795	0.974676	2.83E-02	2758158_G_A	N/A
36	8	95359638	T	G	BMR	0.616264278	7.90E-11	1.91E-11	95594720_C_A	N/A
36	8	95595162	G	A	Body fat %	0.229526936	0.00035829	8.02305E-05	95631555_G_A	N/A
36	8	95595162	G	A	Weight	0.320495945	1.24E-07	3.37E-11	95359638_T_G	N/A
36	8	95595162	G	A	BMI	0.030810207	1.22747E-11	1.22747E-11	95595162_G_A	0.462104964
36	8	95354366	C	G	DBP	0.846217449	0.331449	0.0136669	95595162_G_A	N/A
36	8	95615216	A	G	SBP	0.905307133	0.00455471	0.00341518	95605260_G_A	N/A
36	8	95595162	G	A	HC	0.141421179	1.36E-07	6.09E-11	95359638_T_G	N/A
36	8	95359638	T	G	WC	0.225497233	2.00E-08	2.00E-08	95359638_T_G	N/A
36	8	95530969	G	A	Height	0.548682018	2.93E-01	7.77E-03	95595162_G_A	N/A
36	8	95489281	T	C	PEF	0.943209077	0.0414984	0.0147959	95369109_T_C	N/A
37	8	10939273	G	T	BMR	0.65593528	1.89E-09	8.90E-10	10665444_C_A	N/A
37	8	11281273	A	G	Body fat %	0.05687746	0.00880797	3.63E-07	10665444_C_A	N/A
37	8	11281273	A	G	Weight	0.056437267	1.92E-07	1.54E-11	10665444_C_A	N/A
37	8	11281273	A	G	BMI	0.019184584	3.48932E-13	6.18E-19	10665444_C_A	0.336736787
37	8	10761585	G	A	DBP	0.068478224	1.50991E-14	1.43712E-14	10788875_C_T	N/A
37	8	11024663	C	A	SBP	0.759257952	2.502E-09	9.09793E-14	10604164_T_C	N/A
37	8	11281273	A	G	HC	0.049539799	1.40E-06	2.39E-10	10665444_C_A	0.621762467
37	8	11281273	A	G	WC	0.258655469	0.028632	2.30993E-05	10580550_G_C	N/A
37	8	10580550	G	C	Height	0.92353557	3.53E-01	4.07E-07	10810451_A_G	N/A
37	8	10665444	C	A	PEF	0.851073929	0.000466217	2.85067E-07	10957243_A_G	N/A
38	9	98316094	G	A	BMR	0.01042101	2.44E-19	1.55E-21	98256235_T_G	0.268381152
38	9	98209594	G	A	Body fat %	0.219618083	4.29E-03	4.29E-03	98209594_G_A	N/A
38	9	98316094	G	A	Weight	0.024649646	3.76E-08	5.38E-09	98380222_G_A	0.377428447
38	9	98316331	G	C	BMI	0.251353475	7.23E-01	0.0160085	98316094_G_A	N/A
38	9	98142657	G	A	DBP	0.606178952	0.00442128	0.00442128	98142657_G_A	N/A
38	9	98181368	C	G	SBP	0.885579129	0.355509	0.00215635	98239190_G_C	N/A
38	9	98316094	G	A	HC	0.054065489	1.19E-04	8.11E-06	98185205_G_A	N/A
38	9	98316094	G	A	WC	0.390401484	4.54E-02	1.08E-03	98205443_C_A	N/A
38	9	98185205	G	A	Height	0.990567218	2.50E-29	2.08E-70	98316094_G_A	N/A
38	9	98256235	T	G	PEF	0.906585566	0.00580355	0.00580355	98256235_T_G	N/A
39	9	129464827	C	T	BMR	0.677874862	1.05E-08	1.85E-09	129464856_A_G	N/A
39	9	129464827	C	T	Body fat %	0.029297828	1.44E-03	6.14E-04	129460914_A_G	0.779322664
39	9	129464827	C	T	Weight	0.201782771	1.61E-07	2.95E-08	129460914_A_G	N/A
39	9	129464827	C	T	BMI	0.026155868	5.19806E-08	1.62E-08	129464856_A_G	0.553340265

Locus #	Chrom	Pos	A1	A2	Trait	Lead cFDR P	Assoc P	Lead assoc P	Lead assoc (pos_a1_a2)	p
39	9	129467340	C	T	DBP	0.428446291	0.0263767	0.00964725	129460914_A_G	N/A
39	9	129460914	A	G	SBP	0.959764006	0.0202954	0.0202954	129460914_A_G	N/A
39	9	129464856	A	G	HC	0.103478686	2.02E-07	2.02E-07	129464856_A_G	N/A
39	9	129464827	C	T	WC	0.229674155	4.70E-05	2.10E-05	129464856_A_G	N/A
39	9	129462901	T	C	Height	0.98574984	4.91E-01	3.76E-01	129460914_A_G	N/A
39	9	129467340	C	T	PEF	0.95791195	0.0679593	0.0679593	129467340_C_T	N/A
40	10	104749725	A	G	BMR	0.544786197	1.93E-12	5.52E-14	104610926_G_T	N/A
40	10	104635103	G	A	Body fat %	0.115229232	0.000180365	0.000180365	104635103_G_A	N/A
40	10	104635103	G	A	Weight	0.13162895	3.44E-10	2.19E-12	104610926_G_T	N/A
40	10	104635103	G	A	BMI	0.062096308	1.19503E-08	1.19503E-08	104635103_G_A	N/A
40	10	104680137	T	A	DBP	0.012540984	1.57286E-13	9.48086E-17	104906211_T_C	0.440687101
40	10	104616663	T	C	SBP	0.621354519	6.35E-23	1.41954E-28	104906211_T_C	N/A
40	10	104635103	G	A	HC	0.237288673	2.02666E-09	9.54E-10	104635344_G_C	N/A
40	10	104635103	G	A	WC	0.161909217	4.7412E-06	2.33E-06	104610926_G_T	N/A
40	10	104601565	A	T	Height	0.410629583	2.16E-10	3.00E-33	104278601_T_C	N/A
40	10	104844872	T	C	PEF	0.760525531	8.44E-05	6.01661E-07	104487871_T_C	N/A
41	10	104636655	G	C	BMR	0.724148907	2.80E-08	4.44E-09	104642237_T_G	N/A
41	10	104636655	G	C	Body fat %	0.366038989	7.60E-01	7.60E-01	104636655_G_C	N/A
41	10	104636655	G	C	Weight	0.421937773	8.75E-05	1.44E-05	104642237_T_G	N/A
41	10	104636655	G	C	BMI	0.710162539	0.0790437	5.70E-02	104642237_T_G	N/A
41	10	104642237	T	G	DBP	0.042609177	3.24891E-07	3.24891E-07	104642237_T_G	0.602419518
41	10	104636655	G	C	SBP	1	1.77583E-13	1.20007E-14	104642237_T_G	N/A
41	10	104636655	G	C	HC	0.702816992	9.96E-03	2.72E-03	104642237_T_G	N/A
41	10	104636655	G	C	WC	0.621756064	2.76E-02	1.73E-02	104642237_T_G	N/A
41	10	104636655	G	C	Height	0.973443626	3.24E-08	7.06E-10	104642237_T_G	N/A
41	10	104642237	T	G	PEF	0.766975994	0.00277521	0.00277521	104642237_T_G	N/A
42	11	27583129	T	C	BMR	0.24430835	6.09E-23	1.06E-28	27728102_G_A	N/A
42	11	27456488	T	C	Body fat %	0.006009166	6.67355E-12	1.92E-24	27709630_C_G	0.315200356
42	11	27714884	T	C	Weight	0.0175533	4.11E-35	3.43E-37	27728102_G_A	0.142367245
42	11	27456488	T	C	BMI	0.004652358	3.78748E-17	4.37534E-37	27702383_G_T	0.233369246
42	11	27677586	G	T	DBP	0.277462267	0.144367	0.000984586	27734420_T_C	N/A
42	11	27640223	T	G	SBP	0.874169219	0.428339	0.00327971	27728102_G_A	N/A
42	11	27516785	C	G	HC	0.000976009	4.33E-24	1.38E-33	27728102_G_A	0.280512372
42	11	27456059	C	T	WC	0.000713111	1.6275E-17	1.28E-31	27709630_C_G	0.304525847
42	11	27523186	G	T	Height	0.831709132	6.49E-02	9.52E-03	27703480_T_A	N/A
42	11	27636576	C	T	PEF	0.935301846	0.729971	0.114782	27736207_C_T	N/A
43	11	65245829	T	A	BMR	0.483117431	2.65E-11	2.65E-11	65245829_T_A	N/A
43	11	65294799	C	T	Body fat %	0.091254854	2.28686E-08	2.28686E-08	65294799_C_T	N/A
43	11	65314830	A	T	Weight	0.078517471	6.96E-11	1.26E-11	65245829_T_A	N/A
43	11	65314830	A	T	BMI	0.029576301	2.78708E-09	2.03141E-09	65294799_C_T	0.635310979
43	11	65220206	G	A	DBP	0.899826144	0.560974	0.382412	65245829_T_A	N/A
43	11	65256164	G	A	SBP	0.866152737	0.00182183	0.00182183	65256164_G_A	N/A
43	11	65314830	A	T	HC	0.56480224	4.41E-04	1.42E-04	65245829_T_A	N/A
43	11	65294799	C	T	WC	0.005247311	2.03555E-10	2.03555E-10	65294799_C_T	0.602600835
43	11	65228278	C	T	Height	0.98162151	8.20E-04	2.47E-04	65229828_G_T	N/A
43	11	65229828	G	T	PEF	0.964042442	0.146818	0.103457	65232345_T_C	N/A
44	11	843461	T	C	BMR	0.669967471	6.86E-09	6.81E-09	838842_C_G	N/A
44	11	840363	C	T	Body fat %	0.554900968	7.89E-04	7.89E-04	840363_C_T	N/A
44	11	843461	T	C	Weight	0.351861718	1.07E-07	1.85E-08	840363_C_T	N/A
44	11	843461	T	C	BMI	0.406564968	2.95396E-07	1.77981E-09	840363_C_T	N/A
44	11	840363	C	T	DBP	0.041649087	0.522426	0.213946	843461_T_C	0.998655082
44	11	843461	T	C	SBP	1	0.0316425	0.0316425	843461_T_C	N/A
44	11	840363	C	T	HC	0.424805639	1.45E-08	1.45E-08	840363_C_T	N/A
44	11	843461	T	C	WC	0.547655072	3.65E-04	3.09E-05	840363_C_T	N/A
44	11	840363	C	T	Height	1	7.49E-01	3.03E-01	843461_T_C	N/A
44	11	843461	T	C	PEF	0.954575722	0.0338666	0.03314	840363_C_T	N/A
45	12	50263148	G	A	BMR	0.549161549	9.77E-30	9.77E-30	50263148_G_A	N/A
45	12	50285780	T	G	Body fat %	0.056898308	1.07E-11	1.13E-22	50256063_C_T	N/A
45	12	50285780	T	G	Weight	0.078693615	7.13E-23	1.36E-36	50263148_G_A	N/A
45	12	50285780	T	G	BMI	0.016558793	2.52354E-19	8.27E-31	50263148_G_A	0.317830331
45	12	50205712	G	A	DBP	0.335066507	0.000494527	0.000494527	50205712_G_A	N/A
45	12	50208343	A	G	SBP	0.84973948	0.00141202	0.00129234	50213076_C_T	N/A
45	12	50199316	G	C	HC	0.017825321	3.58998E-13	1.00E-23	50263148_G_A	0.512580375
45	12	50285780	T	G	WC	0.061724656	1.11E-13	2.08E-22	50256063_C_T	N/A



Locus #	Chrom	Pos	A1	A2	Trait	Lead cFDR P	Assoc P	Lead assoc P	Lead assoc (pos_a1_a2)	p
45	12	50278561	T	C	Height	0.986003085	1.23E-01	5.53E-04	50263148_G_A	N/A
45	12	50213254	C	T	PEF	0.959950122	0.653213	0.0254209	50264802_G_A	N/A
46	12	111833788	G	A	BMR	0.104791879	5.23E-18	3.28E-20	112059557_C_T	N/A
46	12	112871372	A	G	Body fat %	0.283852457	2.01E-01	0.00157225	112801608_A_G	N/A
46	12	112486818	A	G	Weight	0.042175981	2.09E-10	1.21E-12	112200150_T_C	0.508460892
46	12	112486818	A	G	BMI	0.354728719	2.23E-05	2.9842E-06	112200150_T_C	N/A
46	12	111973358	A	G	DBP	0.113939055	1.71155E-54	1.71155E-54	111973358_A_G	N/A
46	12	112007756	C	T	SBP	0.185649087	1.16087E-17	2.53624E-18	111884608_T_C	N/A
46	12	112486818	A	G	HC	0.024506181	3.49E-11	2.00E-13	112200150_T_C	0.441977223
46	12	111715197	C	T	WC	0.226516061	3.97E-02	7.56852E-05	112245170_A_C	N/A
46	12	112849899	G	A	Height	0.80359742	1.04E-05	1.39E-10	111884608_T_C	N/A
46	12	112146964	C	T	PEF	0.917147783	0.926677	0.0494916	112179471_T_C	N/A
47	12	120867798	A	G	BMR	0.285753083	5.77E-19	5.77E-19	120867798_A_G	N/A
47	12	120907309	T	A	Body fat %	0.119557449	2.56E-13	2.56E-13	120907309_T_A	N/A
47	12	120907309	T	A	Weight	0.047244877	1.92E-22	1.81E-22	120867798_A_G	0.347775438
47	12	120907309	T	A	BMI	0.090251461	6.4172E-09	6.4172E-09	120907309_T_A	N/A
47	12	120846213	G	A	DBP	0.621447863	0.0147006	0.000507037	120559928_C_T	N/A
47	12	121089313	A	G	SBP	0.891304736	0.415667	0.0606656	120752247_C_T	N/A
47	12	120907309	T	A	HC	0.162153375	1.94E-16	6.63E-17	120867798_A_G	N/A
47	12	120541599	G	A	WC	0.186730402	1.94E-08	1.04E-14	120867798_A_G	N/A
47	12	120384155	G	T	Height	0.975998883	8.91E-12	6.93E-23	120867798_A_G	N/A
47	12	120559928	C	T	PEF	0.953717148	0.0790798	0.0488547	120503552_G_A	N/A
48	12	1035708	T	C	BMR	0.402900513	4.87E-12	5.43E-16	991306_G_A	N/A
48	12	975276	T	G	Body fat %	0.155412844	3.83E-04	4.14E-09	991306_G_A	N/A
48	12	1032472	G	A	Weight	0.026535476	1.10E-09	4.30E-17	991306_G_A	0.548662027
48	12	998365	G	T	BMI	0.046468981	2.25E-07	5.09E-16	991306_G_A	0.600075911
48	12	975276	T	G	DBP	0.885996812	0.728694	0.242116	893613_T_C	N/A
48	12	1028526	C	T	SBP	0.909970413	0.311729	0.00734641	1035708_T_C	N/A
48	12	1032472	G	A	HC	0.021804978	8.00E-06	4.11E-13	991306_G_A	0.703971003
48	12	1022679	A	G	WC	0.029660796	1.43E-05	1.01E-11	991306_G_A	0.712558336
48	12	1034186	G	A	Height	0.560605758	1.71E-04	3.08E-05	1001807_G_A	N/A
48	12	997833	G	A	PEF	0.967204481	0.583582	0.29548	1033183_C_A	N/A
49	12	103761501	G	A	BMR	0.648086367	6.70E-09	2.97E-10	103724090_G_A	N/A
49	12	103761501	G	A	Body fat %	0.179021954	4.30E-07	7.87E-08	103755097_A_G	N/A
49	12	103761501	G	A	Weight	0.21789763	2.20E-11	1.85E-12	103724090_G_A	N/A
49	12	103761501	G	A	BMI	0.042837503	2.65004E-13	1.05E-14	103724090_G_A	0.440924686
49	12	103663536	T	A	DBP	0.801871119	0.00587393	0.00203174	103661104_A_G	N/A
49	12	103661104	A	G	SBP	0.988220397	0.320641	0.308398	103658925_C_G	N/A
49	12	103761501	G	A	HC	0.434114554	3.16E-09	1.21E-11	103706754_C_G	N/A
49	12	103761501	G	A	WC	0.397991334	1.64E-09	9.44E-10	103724090_G_A	N/A
49	12	103720658	C	A	Height	0.983543573	5.42E-01	1.90E-01	103662895_T_C	N/A
49	12	103767871	G	A	PEF	0.960334866	0.943187	0.510004	103724090_G_A	N/A
50	13	97081460	T	G	BMR	0.259589692	1.74E-08	3.17E-10	96922449_G_A	N/A
50	13	97060551	T	C	Body fat %	0.187595285	1.31E-04	1.74E-06	97047020_G_A	N/A
50	13	97081460	T	G	Weight	0.022310831	7.32E-09	5.29E-12	97047020_G_A	0.727548773
50	13	97020705	C	G	BMI	0.192015542	1.05E-11	4.14621E-12	97047020_G_A	N/A
50	13	96922449	G	A	DBP	0.757796417	0.0569201	0.0409497	97034410_G_A	N/A
50	13	97201221	C	T	SBP	0.891388356	0.306014	0.000100702	97034410_G_A	N/A
50	13	97121087	G	A	HC	0.034404021	8.98032E-07	6.06E-10	97047020_G_A	0.876383291
50	13	97121087	G	A	WC	0.050803063	1.17404E-06	9.98E-10	97047020_G_A	N/A
50	13	97085948	A	G	Height	0.480990376	1.23E-04	6.71E-05	97081460_T_G	N/A
50	13	96928767	T	C	PEF	0.959152035	0.747903	0.496629	96922191_T_C	N/A
51	14	25948867	C	T	BMR	0.422238944	1.14E-19	2.46E-20	25927832_A_G	N/A
51	14	25965098	T	C	Body fat %	0.642811501	0.0122699	1.34E-05	25930988_C_A	N/A
51	14	25948867	C	T	Weight	0.201666622	5.74E-16	2.34E-17	25927832_A_G	N/A
51	14	25949019	T	G	BMI	0.279176293	2.45711E-15	2.44854E-16	25930988_C_A	N/A
51	14	25946258	C	T	DBP	0.859336424	0.675854	0.337258	25925165_T_A	N/A
51	14	25956292	G	A	SBP	0.993952059	0.599028	0.434895	25946258_C_T	N/A
51	14	25950864	C	T	HC	0.125594231	2.04E-08	5.36E-13	25932585_A_G	N/A
51	14	25933965	G	T	WC	0.049518435	1.51E-09	4.79E-10	25935161_G_A	0.461125426
51	14	25944306	C	A	Height	0.985915985	9.54E-02	3.53E-02	25925165_T_A	N/A
51	14	25946258	C	T	PEF	0.94139405	0.800435	0.328555	25925403_C_T	N/A
52	14	103380403	G	T	BMR	0.599178619	3.76E-09	3.70E-11	103246470_A_G	N/A
52	14	103380403	G	T	Body fat %	0.254222707	1.19E-10	8.42E-12	103246470_A_G	N/A

Locus #	Chrom	Pos	A1	A2	Trait	Lead cFDR P	Assoc P	Lead assoc P	Lead assoc (pos_a1_a2)	p
52	14	103380403	G	T	Weight	0.18022383	4.15E-13	2.13E-15	103246470_A_G	N/A
52	14	103380403	G	T	BMI	0.20101626	3.697E-14	3.08E-16	103246470_A_G	N/A
52	14	103326780	T	C	DBP	0.832835359	0.00827298	0.000257937	103249127_G_C	N/A
52	14	103326780	T	C	SBP	0.883413862	0.480493	0.260119	103387971_C_G	N/A
52	14	103360000	A	G	HC	0.316829079	2.74E-09	5.12E-12	103246470_A_G	N/A
52	14	103380403	G	T	WC	0.015919514	6.28E-13	2.75E-13	103246470_A_G	0.430738993
52	14	103256961	A	G	Height	0.978779168	9.46E-01	6.68E-01	103246470_A_G	N/A
52	14	103256199	C	G	PEF	0.960776136	0.683238	1.16E-01	103269755_C_A	N/A
53	14	29681294	G	A	BMR	0.54623277	5.59E-11	4.39E-11	29681138_G_C	N/A
53	14	29681138	G	C	Body fat %	0.034814319	2.13E-11	2.13E-11	29681138_G_C	0.461999574
53	14	29677727	T	C	Weight	0.18720059	1.77E-12	3.64E-14	29680331_G_A	N/A
53	14	29677727	T	C	BMI	0.175350161	4.85969E-11	3.27E-13	29681138_G_C	N/A
53	14	29702590	T	G	DBP	0.852609282	0.645917	0.311985	29726942_C_A	N/A
53	14	29684521	T	C	SBP	0.883575821	0.441895	0.196457	29681138_G_C	N/A
53	14	29677727	T	C	HC	0.32167856	1.08E-12	1.29E-13	29681138_G_C	N/A
53	14	29677727	T	C	WC	0.072406231	8.24E-11	7.21E-12	29680331_G_A	N/A
53	14	29729003	G	A	Height	0.985584671	3.81E-01	3.61E-02	29681294_G_A	N/A
53	14	29677727	T	C	PEF	0.955460135	0.941756	0.477354	29727517_G_A	N/A
54	14	94028197	C	T	BMR	0.505340789	8.26E-09	4.82E-11	94073093_A_G	N/A
54	14	94073093	A	G	Body fat %	0.016176104	3.97E-08	1.13E-12	94023972_G_A	0.47162039
54	14	94028197	C	T	Weight	0.030449004	7.10E-14	1.72E-15	94023972_G_A	0.390102385
54	14	94028197	C	T	BMI	0.043712251	6.60677E-19	6.48E-20	94023972_G_A	0.231960483
54	14	94023972	G	A	DBP	0.754385345	0.311833	0.0665612	93820274_C_T	N/A
54	14	94135104	T	C	SBP	0.901885167	2.73E-01	0.0335752	93702672_T_A	N/A
54	14	94054707	T	C	HC	0.028061688	4.22776E-13	3.60E-16	94023972_G_A	0.293969388
54	14	94007075	C	T	WC	0.01439818	1.80566E-14	5.77E-15	94023972_G_A	0.284131562
54	14	94135104	T	C	Height	0.891895665	9.57E-01	1.99E-02	93912668_C_A	N/A
54	14	94031914	A	G	PEF	0.89855997	0.0578662	0.0285052	93715988_C_G	N/A
55	14	101531854	A	G	BMR	0.602876133	7.03E-09	7.03E-09	101531854_A_G	N/A
55	14	101531854	A	G	Body fat %	0.024541258	8.90E-05	8.90E-05	101531854_A_G	0.603207328
55	14	101531854	A	G	Weight	0.061451862	5.70E-09	5.70E-09	101531854_A_G	N/A
55	14	101531854	A	G	BMI	0.041214793	2.80E-09	2.80E-09	101531854_A_G	0.422554231
55	14	101531854	A	G	DBP	0.909186686	0.346215	0.346215	101531854_A_G	N/A
55	14	101531854	A	G	SBP	0.989274841	0.145421	0.145421	101531854_A_G	N/A
55	14	101531854	A	G	HC	0.201996842	5.59E-08	5.59E-08	101531854_A_G	N/A
55	14	101531854	A	G	WC	0.17713209	8.54E-09	8.54E-09	101531854_A_G	N/A
55	14	101531854	A	G	Height	1	6.97E-01	6.97E-01	101531854_A_G	N/A
55	14	101531854	A	G	PEF	0.97669192	0.663229	0.663229	101531854_A_G	N/A
56	15	99196678	C	T	BMR	0.197501081	1.11E-09	5.13E-23	99186488_G_C	N/A
56	15	99196678	C	T	Body fat %	0.246151041	5.54E-03	0.00104698	99210597_G_C	N/A
56	15	99196678	C	T	Weight	0.011276677	2.50E-03	5.03E-10	99186488_G_C	0.847938607
56	15	99196678	C	T	BMI	0.111544338	0.021358	0.00680232	99210597_G_C	N/A
56	15	99196678	C	T	DBP	0.392499969	0.0226466	0.00662338	99180380_A_C	N/A
56	15	99186250	G	A	SBP	0.895496174	0.645121	0.00468427	99192535_C_T	N/A
56	15	99196678	C	T	HC	0.00639994	2.94E-01	6.14E-07	99183876_T_C	0.994218607
56	15	99196678	C	T	WC	0.046206465	4.01E-01	0.150711	99210597_G_C	0.987916732
56	15	99193276	A	C	Height	0.359673902	2.74E-26	3.47E-68	99194896_C_G	N/A
56	15	99183876	T	C	PEF	0.891342914	8.96793E-05	9.69802E-06	99190601_G_A	N/A
57	16	53802494	C	T	BMR	2.26683E-08	7.97E-138	7.97E-138	53802494_C_T	0.162900473
57	16	53802494	C	T	Body fat %	2.37807E-08	3.10E-96	3.10E-96	53802494_C_T	0.201104781
57	16	53802494	C	T	Weight	7.52776E-15	7.69E-170	7.69E-170	53802494_C_T	0.170159273
57	16	53802494	C	T	BMI	8.09319E-13	6.39E-192	6.39E-192	53802494_C_T	0.126505341
57	16	53834607	T	C	DBP	0.210525155	0.0762335	0.00233064	53798523_G_A	N/A
57	16	53800387	A	G	SBP	0.635892508	0.0056188	7.58945E-06	53807764_A_G	N/A
57	16	53806453	A	G	HC	9.8432E-11	5.27E-137	5.27E-137	53806453_A_G	0.159912361
57	16	53802494	C	T	WC	1.07784E-14	1.41E-145	1.41E-145	53802494_C_T	0.176452164
57	16	53798622	G	T	Height	0.677591436	5.59E-01	1.54E-02	53806145_T_C	N/A
57	16	53824226	A	G	PEF	0.928217585	0.706211	0.0304904	53755146_A_G	N/A
58	16	30125840	G	C	BMR	0.215356118	7.61E-32	7.43E-52	30045789_C_G	N/A
58	16	30097630	C	T	Body fat %	0.020196953	1.22E-07	2.37E-16	29926552_T_C	0.529981164
58	16	30134656	T	C	Weight	0.036965928	1.02E-33	4.85E-48	29994922_C_T	0.193610163
58	16	30376691	G	A	BMI	0.033676911	3.11311E-08	1.68359E-33	29954654_T_G	0.594208542
58	16	29885698	A	G	DBP	0.6770583	4.96E-03	0.00192427	29958216_G_A	N/A
58	16	29885698	A	G	SBP	0.881645336	0.000463528	0.000422663	29958216_G_A	N/A

Locus #	Chrom	Pos	A1	A2	Trait	Lead cFDR P	Assoc P	Lead assoc P	Lead assoc (pos_a1_a2)	p
58	16	30139900	T	G	HC	0.022090769	2.41E-20	2.09647E-34	29954654_T_G	0.299605154
58	16	29994922	C	T	WC	0.025559914	2.09E-38	2.09E-38	29994922_C_T	0.112894578
58	16	29932691	T	C	Height	0.956757922	5.58E-10	7.36E-22	30048553_C_T	N/A
58	16	30068469	T	G	PEF	0.914695284	0.0257903	1.38E-02	29963812_C_T	N/A
59	16	4014282	C	G	BMR	0.121676437	1.29E-30	1.29E-30	4014282_C_G	N/A
59	16	4017392	T	A	Body fat %	0.001749538	5.71E-14	6.80E-18	4015729_C_T	0.419140003
59	16	4015729	C	T	Weight	0.001488005	4.97E-33	4.97E-33	4015729_C_T	0.219634964
59	16	4015729	C	T	BMI	0.001533762	4.93097E-25	4.93097E-25	4015729_C_T	0.238978631
59	16	4003974	T	C	DBP	0.255114235	0.00278263	2.68972E-08	4014282_C_G	N/A
59	16	4006163	A	T	SBP	0.312048617	0.0250987	0.00293857	4014282_C_G	N/A
59	16	4015729	C	T	HC	0.001174165	1.34E-27	1.34E-27	4015729_C_T	0.235748409
59	16	4015729	C	T	WC	0.002117565	3.13E-21	3.13E-21	4015729_C_T	0.297050615
59	16	4035068	A	G	Height	0.758227765	0.00152439	1.40E-15	4015313_A_G	N/A
59	16	4002320	C	G	PEF	0.921040088	0.0201169	0.0186105	4003974_T_C	N/A
60	16	28944396	C	G	BMR	0.423262324	2.81E-17	7.68E-21	28649651_C_A	N/A
60	16	28542172	T	C	Body fat %	0.007313628	4.06454E-20	3.37E-35	28868962_C_G	0.292462563
60	16	28542172	T	C	Weight	0.018458164	5.79E-21	2.48E-35	28504181_G_A	0.259414603
60	16	28347140	A	C	BMI	0.027930549	1.86E-26	1.71E-27	28649651_C_A	0.192820833
60	16	28336882	T	C	DBP	0.631459876	0.0423541	0.000550524	28532188_T_G	N/A
60	16	28582849	A	G	SBP	0.857376802	0.00147017	0.00147017	28582849_A_G	N/A
60	16	28542172	T	C	HC	0.014381264	1.18289E-22	2.09E-39	28868962_C_G	0.242955409
60	16	28542172	T	C	WC	0.008803196	6.23685E-19	9.09927E-34	28832382_C_T	0.23950203
60	16	28871191	C	G	Height	0.695370818	2.98E-06	2.82E-07	28718755_A_G	N/A
60	16	28977020	G	A	PEF	0.725401282	0.134479	0.0792737	28937259_T_C	N/A
61	16	31165795	G	A	BMR	0.436878028	1.07E-09	9.23E-15	31025641_C_T	N/A
61	16	30820866	T	C	Body fat %	0.122022139	5.43993E-13	3.34E-17	31025641_C_T	N/A
61	16	31054040	G	C	Weight	0.149451607	4.12E-21	2.12E-22	31025641_C_T	N/A
61	16	31336719	G	A	BMI	0.14410213	3.38234E-12	7.25E-25	31025641_C_T	N/A
61	16	31105554	A	C	DBP	0.329825172	2.91E-07	4.84E-11	30585535_T_C	N/A
61	16	30833246	C	T	SBP	0.575195383	0.00293246	0.000309817	31118024_C_A	N/A
61	16	31011821	T	G	HC	0.253313906	3.44E-27	3.14E-27	31011183_A_G	N/A
61	16	31014179	C	T	WC	0.007330149	1.10E-21	3.37E-22	31025641_C_T	0.254077401
61	16	31185882	G	A	Height	0.982463146	0.730283	1.24E-03	30823047_G_T	N/A
61	16	31229022	G	A	PEF	0.923495006	0.8663	0.0683319	31363788_C_T	N/A
62	17	76665365	A	G	BMR	0.613588923	1.52E-10	4.52E-13	76746325_G_A	N/A
62	17	76746325	G	A	Body fat %	0.249544411	6.48E-04	1.51E-04	76799795_G_A	N/A
62	17	76685106	A	G	Weight	0.293406541	8.94E-09	1.02E-11	76739141_G_A	N/A
62	17	76831800	T	C	BMI	0.378863158	0.00307636	0.000152207	76739141_G_A	N/A
62	17	76739850	C	T	DBP	0.015867251	3.9583E-08	2.10035E-08	76696986_G_A	0.621427804
62	17	76661528	G	A	SBP	0.758837738	6.29093E-09	1.80563E-10	76798362_C_T	N/A
62	17	76684970	G	A	HC	0.405281407	1.80E-04	9.16E-07	76739141_G_A	N/A
62	17	76828154	A	G	WC	0.331472537	0.000951906	1.07E-05	76739141_G_A	N/A
62	17	76789991	A	T	Height	0.670607018	6.64E-14	1.33E-14	76790279_C_T	N/A
62	17	76685106	A	G	PEF	0.88288213	0.00754951	0.00488715	76718842_C_T	N/A
63	17	1820080	G	A	BMR	0.669898802	6.81E-09	1.03E-10	1844519_G_A	N/A
63	17	1849662	G	A	Body fat %	0.2624483	1.88E-10	7.52E-20	1835482_C_T	N/A
63	17	1820080	G	A	Weight	0.299185608	5.93E-15	1.35E-17	1835482_C_T	N/A
63	17	1835482	C	T	BMI	0.286485295	7.49E-19	7.49E-19	1835482_C_T	N/A
63	17	1866892	G	T	DBP	0.337285986	0.0161086	5.95805E-05	1849663_C_A	N/A
63	17	1849662	G	A	SBP	0.844473214	0.000290258	0.000289966	1849663_C_A	N/A
63	17	1820080	G	A	HC	0.376156244	3.23E-15	8.31E-17	1824305_C_A	N/A
63	17	1820080	G	A	WC	0.045633399	1.52E-14	3.11E-17	1835482_C_T	0.478951977
63	17	1846831	C	A	Height	0.992620753	7.82E-01	2.89E-01	1866892_G_T	N/A
63	17	1830836	C	T	PEF	0.827120094	0.00059473	1.49325E-05	1853400_G_C	N/A
64	17	65800140	G	A	BMR	0.460479953	1.19E-08	2.38E-09	65832016_T_C	N/A
64	17	65828371	T	A	Body fat %	0.013819782	3.84E-22	3.58E-23	65836001_A_C	0.351807448
64	17	66004689	T	G	Weight	0.048111668	2.30E-14	4.37E-19	65837235_A_G	0.30366108
64	17	65828371	T	A	BMI	0.192974953	1.23E-10	8.05E-12	65832016_T_C	N/A
64	17	65892343	G	T	DBP	0.630045246	3.57E-01	0.136709	65800140_G_A	N/A
64	17	65892064	T	A	SBP	0.877069112	0.424261	0.0393061	65960854_T_C	N/A
64	17	65885911	C	T	HC	0.149247527	1.12E-10	4.69E-15	65837235_A_G	N/A
64	17	65892507	G	C	WC	0.033720933	1.21224E-17	2.20E-20	65836001_A_C	0.282283182
64	17	65968003	A	G	Height	0.957257245	3.31E-08	9.82E-09	65892343_G_T	N/A
64	17	65977053	C	T	PEF	0.936492097	0.655713	0.204969	66007279_T_C	N/A

Locus #	Chrom	Pos	A1	A2	Trait	Lead cFDR P	Assoc P	Lead assoc P	Lead assoc (pos_a1_a2)	p
65	18	57829135	T	C	BMR	0.002912981	7.12E-131	7.12E-131	57829135_T_C	0.046781485
65	18	57829135	T	C	Body fat %	1.33628E-06	1.90E-25	1.13E-25	57850422_G_T	0.33974902
65	18	57829135	T	C	Weight	1.36714E-06	7.15E-114	7.15E-114	57829135_T_C	0.091566496
65	18	57829135	T	C	BMI	6.60627E-07	2.54E-75	1.54964E-75	57848651_A_G	0.14602229
65	18	58053203	C	G	DBP	0.517341605	0.0145459	0.000271105	57869750_C_T	N/A
65	18	57846077	C	T	SBP	0.88516826	0.517474	0.0107927	58022740_G_C	N/A
65	18	57802714	G	A	HC	2.74094E-07	1.04E-56	1.35E-68	57829135_T_C	0.214206749
65	18	57829135	T	C	WC	1.22375E-07	1.86E-65	1.45E-65	57848651_A_G	0.167059809
65	18	57802714	G	A	Height	0.298231216	1.54E-26	6.85E-30	57829135_T_C	N/A
65	18	57890756	G	T	PEF	0.901203295	0.0301395	0.00160984	58030066_T_A	N/A
66	18	57955945	T	G	BMR	0.300054428	1.07E-33	1.03E-38	57958244_C_T	N/A
66	18	57997729	C	T	Body fat %	0.024651119	5.66E-17	2.57E-19	57961249_C_T	0.250035583
66	18	57967655	A	C	Weight	0.007034753	2.64E-19	5.81E-43	57961249_C_T	0.258213459
66	18	57967655	A	C	BMI	0.017810572	3.47868E-17	3.39E-38	57961249_C_T	0.236602448
66	18	57985366	T	C	DBP	0.389048539	0.00375156	0.000436143	57981916_C_A	N/A
66	18	57981916	C	A	SBP	0.94111167	0.00908446	0.00908446	57981916_C_A	N/A
66	18	57967655	A	C	HC	0.022238492	4.00E-16	1.18E-26	57961249_C_T	0.272323883
66	18	57967655	A	C	WC	0.000114937	1.13759E-15	1.60E-31	57961249_C_T	0.382552715
66	18	57948098	T	C	Height	0.563062706	0.295555	4.97E-05	57958244_C_T	N/A
66	18	57921433	A	T	PEF	0.898239598	1.25E-02	0.00990152	57934238_T_C	N/A
67	18	21137442	A	G	BMR	0.55817787	3.29E-13	1.66E-21	21109466_G_T	N/A
67	18	21137442	A	G	Body fat %	0.036357684	7.5203E-12	4.42091E-16	21080859_C_A	0.611122101
67	18	21137442	A	G	Weight	0.056487593	1.23E-16	1.32E-23	21109466_G_T	N/A
67	18	21137442	A	G	BMI	0.284076644	1.70808E-10	2.67458E-14	21090023_A_G	N/A
67	18	21003267	C	G	DBP	0.131369497	0.00440781	7.30293E-05	20897797_C_G	N/A
67	18	21133937	G	A	SBP	0.892748676	3.42E-01	0.0732569	21109466_G_T	N/A
67	18	21137442	A	G	HC	0.093462043	1.97548E-14	1.36E-18	21126952_C_T	N/A
67	18	21137442	A	G	WC	0.095982894	1.27677E-12	1.58E-19	21109466_G_T	N/A
67	18	21074922	G	A	Height	0.947142117	2.85E-06	1.96E-20	20890718_G_T	N/A
67	18	21074922	G	A	PEF	0.922434386	0.663676	0.0505057	21069137_G_T	N/A
68	18	57913703	T	C	BMR	0.436448949	2.34E-18	2.34E-18	57913703_T_C	N/A
68	18	57805566	G	A	Body fat %	0.576550027	5.81E-04	6.28E-05	57913703_T_C	N/A
68	18	57913703	T	C	Weight	0.056229522	1.35E-17	1.35E-17	57913703_T_C	N/A
68	18	57913703	T	C	BMI	0.093372719	6.08E-13	6.08E-13	57913703_T_C	N/A
68	18	57913703	T	C	DBP	0.543306852	0.0583561	0.0583561	57913703_T_C	N/A
68	18	57902986	T	C	SBP	0.888743541	0.571271	0.414622	57913703_T_C	N/A
68	18	57913703	T	C	HC	0.262508484	2.86E-10	2.86E-10	57913703_T_C	N/A
68	18	57913703	T	C	WC	0.013694334	1.08E-12	1.08E-12	57913703_T_C	0.411437602
68	18	57948257	T	C	Height	0.984357831	8.08E-04	2.05E-05	57863787_A_G	N/A
68	18	57948257	T	C	PEF	0.928728335	0.0287188	0.0277088	57965832_G_A	N/A
69	18	57809744	C	T	BMR	0.821673947	1.16E-17	1.16E-17	57809744_C_T	N/A
69	18	57809744	C	T	Body fat %	0.117718759	1.10E-05	1.10E-05	57809744_C_T	N/A
69	18	57809744	C	T	Weight	0.300871343	7.89E-16	7.89E-16	57809744_C_T	N/A
69	18	57809744	C	T	BMI	0.17004344	7.11E-13	7.11E-13	57809744_C_T	N/A
69	18	57809744	C	T	DBP	0.944323754	0.496845	0.496845	57809744_C_T	N/A
69	18	57809744	C	T	SBP	1	0.964222	0.964222	57809744_C_T	N/A
69	18	57809744	C	T	HC	0.399705353	8.33E-11	8.33E-11	57809744_C_T	N/A
69	18	57809744	C	T	WC	0.038868995	4.71E-12	4.71E-12	57809744_C_T	0.391633563
69	18	57809744	C	T	Height	1	1.80E-02	1.80E-02	57809744_C_T	N/A
69	18	57809744	C	T	PEF	0.979981602	0.963139	0.963139	57809744_C_T	N/A
70	19	30266706	G	A	BMR	0.332909589	3.84E-20	7.17E-22	30290357_G_C	N/A
70	19	30265235	G	A	Body fat %	0.052192771	1.88E-12	1.66E-13	30294991_G_T	N/A
70	19	30265235	G	A	Weight	0.009380638	1.38E-21	1.28E-23	30290357_G_C	0.273402478
70	19	30266706	G	A	BMI	0.00785281	4.83268E-15	7.72E-16	30290357_G_C	0.415999842
70	19	30255090	G	A	DBP	0.113670143	9.33278E-11	4.7535E-19	30288177_G_A	N/A
70	19	30249397	C	G	SBP	0.804252296	0.000777643	5.8359E-05	30305684_G_A	N/A
70	19	30265235	G	A	HC	0.027412606	1.78E-13	4.19E-14	30305684_G_A	0.334001762
70	19	30265235	G	A	WC	0.029727296	2.33E-16	1.24E-17	30290357_G_C	0.367318415
70	19	30343667	G	C	Height	0.966632317	1.17E-07	6.15E-09	30305684_G_A	N/A
70	19	30313603	C	G	PEF	0.884406614	0.00333688	0.00197873	30300841_G_A	N/A
71	19	2176403	G	A	BMR	0.675924313	3.67E-18	3.67E-18	2176403_G_A	N/A
71	19	2198619	C	T	Body fat %	0.723076299	0.732931	0.015334	2192015_C_T	N/A
71	19	2192634	G	A	Weight	0.382675737	2.58E-08	2.46E-10	2186757_T_C	N/A
71	19	2195065	G	T	BMI	0.660003125	0.942624	0.0206876	2186757_T_C	N/A

Locus #	Chrom	Pos	A1	A2	Trait	Lead cFDR P	Assoc P	Lead assoc P	Lead assoc (pos_a1_a2)	p
71	19	2205668	C	G	DBP	0.044505536	1.22053E-06	3.90E-07	2198619_C_T	0.731396717
71	19	2207041	A	G	SBP	0.882039416	0.487541	9.68827E-05	2197734_G_A	N/A
71	19	2192634	G	A	HC	0.45031647	5.25E-08	1.19E-10	2186757_T_C	N/A
71	19	2145450	T	C	WC	0.538633526	0.793858	4.87E-02	2186757_T_C	N/A
71	19	2199351	T	C	Height	0.338975574	3.31438E-26	1.78E-58	2155042_G_A	N/A
71	19	2191980	T	C	PEF	0.945213812	0.574725	0.0821891	2156954_T_C	N/A
72	20	33954913	C	A	BMR	0.159097391	5.80E-75	2.31E-111	34025756_A_G	N/A
72	20	33213196	A	C	Body fat %	0.345747346	0.899691	0.000185875	34145773_C_T	N/A
72	20	34158394	A	G	Weight	0.185064208	5.2484E-19	1.93E-53	34025756_A_G	N/A
72	20	34154425	T	C	BMI	0.331483984	0.0517891	1.80E-04	33844938_A_C	N/A
72	20	33744323	G	C	DBP	0.250582369	0.811229	7.43361E-05	34712310_T_C	N/A
72	20	33750479	T	C	SBP	0.804006775	0.000936799	0.000152688	33880363_A_T	N/A
72	20	33280836	C	T	HC	0.123609821	1.64598E-07	1.60E-30	34025756_A_G	N/A
72	20	34030606	G	A	WC	0.154091691	3.68592E-06	6.03983E-07	34375497_G_A	N/A
72	20	34025756	A	G	Height	0.042109879	8.85E-292	8.85E-292	34025756_A_G	0.019327587
72	20	34185161	G	T	PEF	0.78172655	0.000101738	2.04414E-19	34025756_A_G	N/A
73	20	32077925	C	T	BMR	0.426049364	3.18E-09	4.14E-31	32300634_A_C	N/A
73	20	32026350	C	G	Body fat %	0.022910021	0.239543	0.00508457	32019417_T_C	0.972627687
73	20	32254831	G	T	Weight	0.032128724	5.04E-04	5.31E-19	32289763_A_G	0.453201617
73	20	32245869	C	G	BMI	0.198466659	0.950966	9.81E-05	32505658_G_C	N/A
73	20	32365469	T	C	DBP	0.582007581	0.00599631	9.63135E-06	32298286_C_T	N/A
73	20	32325270	G	A	SBP	0.891924552	0.602071	0.058945	32327399_T_C	N/A
73	20	32245869	C	G	HC	0.135723084	1.15E-02	1.49E-14	32333181_G_T	N/A
73	20	32026350	C	G	WC	0.118464709	0.260485	5.35E-05	32505658_G_C	N/A
73	20	32199872	G	A	Height	0.35112472	6.25E-15	6.49E-54	32304653_C_G	N/A
73	20	32358788	C	T	PEF	0.861774352	0.000288037	7.30457E-05	32188142_C_T	N/A
74	22	42038786	C	T	BMR	0.646371909	1.16E-10	2.31E-19	42070374_A_C	N/A
74	22	42038786	C	T	Body fat %	0.467272287	5.59E-01	1.11E-06	41884954_A_G	N/A
74	22	42287203	C	G	Weight	0.651812058	1.60E-05	6.50E-13	42070374_A_C	N/A
74	22	42287203	C	G	BMI	0.613620123	3.50E-03	3.17682E-08	41804716_G_A	N/A
74	22	42248860	G	A	DBP	0.048818986	2.14856E-05	2.14856E-05	42248860_G_A	0.601719508
74	22	42339525	G	A	SBP	0.939131496	0.485512	0.0166884	42269628_C_T	N/A
74	22	42287203	C	G	HC	0.739308528	3.92E-02	2.40E-06	41884954_A_G	N/A
74	22	41804716	G	A	WC	0.570642786	1.04E-06	1.04E-06	41804716_G_A	N/A
74	22	42339525	G	A	Height	0.974556542	1.27E-10	2.69E-16	42184143_C_T	N/A
74	22	42287203	C	G	PEF	0.859508417	0.000317339	1.05325E-05	42070374_A_C	N/A

**Table B2 Overlap of loci containing significant interaction effects for the ten considered phenotypes** | Tables summarising the number of the 74 loci (column 2) with significant interaction effects (cFDR adjusted  $P < 0.05$ , as described in Section 4.3.4) that are unique to any given trait (column 1; single trait loci), shared across at least two, three, four, five and six traits (column 1, no loci are shared amongst more than six traits).

Single trait loci	
Trait	Number of loci
DBP	8
WC	8
Body fat %	5
BMI	5
Weight	3
HC	3
Height	1
BMR	0

Loci shared across at least two traits	
Traits	Number of loci
Weight, HC	23
Weight, WC	23
BMI, HC	21
HC, WC	21
BMI, WC	18
Weight, BMI	17
Body fat %, Weight	15
Body fat %, WC	14
Body fat %, BMI	13
Body fat %, HC	12
BMR, Weight	7
BMR, WC	6
BMR, BMI	5
BMR, HC	5
DBP, WC	5
BMR, Body fat %	4
Weight, DBP	4
BMI, DBP	3
DBP, HC	3
Body fat %, DBP	2
BMR, DBP	1

Loci shared across at least three traits	
Traits	Number of loci
Weight, HC, WC	19
Weight, BMI, HC	17
BMI, HC, WC	17
Weight, BMI, WC	16
Body fat %, Weight, WC	14
Body fat %, Weight, HC	12
Body fat %, HC, WC	12
Body fat %, Weight, BMI	11
Body fat %, BMI, HC	11
Body fat %, BMI, WC	11
BMR, Weight, BMI	5
BMR, Weight, HC	5
BMR, Weight, WC	5
BMR, BMI, HC	5
BMR, BMI, WC	5
BMR, HC, WC	5
BMR, Body fat %, Weight	4
BMR, Body fat %, BMI	4
BMR, Body fat %, HC	4

Loci shared across at least four traits	
Traits	Number of loci
Weight, BMI, HC, WC	16
Body fat %, Weight, HC, WC	12
Body fat %, Weight, BMI, HC	11
Body fat %, Weight, BMI, WC	11
Body fat %, BMI, HC, WC	11
BMR, Weight, BMI, HC	5
BMR, Weight, BMI, WC	5
BMR, Weight, HC, WC	5
BMR, BMI, HC, WC	5
BMR, Body fat %, Weight, BMI	4
BMR, Body fat %, Weight, HC	4
BMR, Body fat %, Weight, WC	4
BMR, Body fat %, BMI, HC	4
BMR, Body fat %, BMI, WC	4
BMR, Body fat %, HC, WC	4
Weight, DBP, HC, WC	3
Body fat %, Weight, DBP, HC	2
Body fat %, Weight, DBP, WC	2
Body fat %, DBP, HC, WC	2
Weight, BMI, DBP, HC	2
Weight, BMI, DBP, WC	2
BMI, DBP, HC, WC	2
BMR, Weight, BMI, DBP	1
BMR, Weight, DBP, HC	1
BMR, Weight, DBP, WC	1
BMR, BMI, DBP, HC	1
BMR, BMI, DBP, WC	1
BMR, DBP, HC, WC	1
Body fat %, Weight, BMI, DBP	1
Body fat %, BMI, DBP, HC	1
Body fat %, BMI, DBP, WC	1

Loci shared across at least five traits	
Traits	Number of loci
Body fat %, Weight, BMI, HC, WC	11
BMR, Weight, BMI, HC, WC	5
BMR, Body fat %, Weight, BMI, HC	4
BMR, Body fat %, Weight, BMI, WC	4
BMR, Body fat %, Weight, HC, WC	4
BMR, Body fat %, BMI, HC, WC	4
Body fat %, Weight, DBP, HC, WC	2
Weight, BMI, DBP, HC, WC	2
BMR, Weight, BMI, DBP, HC	1
BMR, Weight, BMI, DBP, WC	1
BMR, Weight, DBP, HC, WC	1
BMR, BMI, DBP, HC, WC	1
Body fat %, Weight, BMI, DBP, HC	1
Body fat %, Weight, BMI, DBP, WC	1
Body fat %, BMI, DBP, HC, WC	1

Loci shared across at least six traits	
Traits	Number of loci
BMR, Body fat %, Weight, BMI, HC, WC	4
BMR, Weight, BMI, DBP, HC, WC	1

<b>Traits</b>	<b>Number of loci</b>
BMR, Body fat %, WC	4
Weight, DBP, WC	4
Weight, DBP, HC	3
DBP, HC, WC	3
Body fat %, Weight, DBP	2
Body fat %, DBP, HC	2
Body fat %, DBP, WC	2
Weight, BMI, DBP	2
BMI, DBP, HC	2
BMI, DBP, WC	2
BMR, Weight, DBP	1
BMR, BMI, DBP	1
BMR, DBP, HC	1
BMR, DBP, WC	1
Body fat %, BMI, DBP	1

<b>Traits</b>	<b>Number of loci</b>
Body fat %, Weight, BMI, DBP, HC, WC	1



